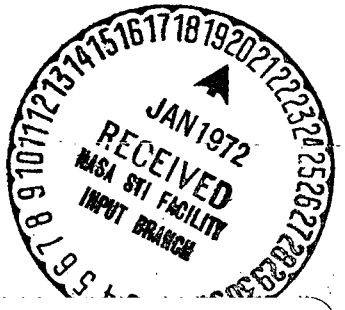
STABILIZATION OF OPTIMAL DIFFERENCE PROCEDURES FOR NUMERICAL INTEGRATION OF ORDINARY DIFFERENTIAL EQUATIONS

H. Brunner

Translation of "Stabilisierung optimaler differenzenverfahren zur numerischen Integration gewöhnlicher Differentialgleichungen"

Doctoral Dissertation No. 4182, Technische Hochschule Zurich, Juris Durck + Verlag, Zurich, 1969, 86 pages



N72-14627 (NASA-TT-F-13856) STABILIZATION OF OPTIMAL DIFFERENCE PROCEDURES FOR NUMERICAL INTEGRATION OF ORDINARY DIFFERENTIAL
Unclas EQUATIONS H. Brunner (Scripta Technica, Inc.) Jun. 1971 72 p CSCL 12A G3/19
12704

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
WASHINGTON, D. C. 20546 JUNE 1971

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
U S Department of Commerce
Springfield VA 22151

The author thanks the National Research Council of Canada for its Contribution, by means of grant number A4805, to the printing expenses of this dissertation.

TABLE OF CONTENTS

Chapter 1.	Introduction	1
Chapter 2.	Modified Optimal Difference Procedures	13
Chapter 3.	Stabilization	32
Chapter 4.	Systems of Differential Equations	42
Chapter 5.	Conditional Stability in the Case of an Explicit Procedure of Maximum Order	51
Chapter 6.	Optimal Procedures for $k = 2$, $k = 4$, and $k = 6$	54
Chapter 7.	Modified Explicit Procedures for $k = 2$, $k = 4$ and $k = 8$	58
Chapter 8.	Examples	62
References		67

STABILIZATION OF OPTIMAL DIFFERENCE PROCEDURES FOR NUMERICAL INTEGRATION OF ORDINARY DIFFERENTIAL EQUATIONS

H. Brunner

The author presents an exposition of new procedures for approximation of particular solutions of linear and nonlinear differential equations (or systems of equations) by means of difference equations. The results of the classical and modified procedures are compared in the case of specific examples. Questions of convergence, stability (relative and absolute), consistency, and optimality are treated.

CHAPTER 1

INTRODUCTION

/7*

1.1. Posing of the problem.

Consider a first-order ordinary differential equation

$$x' = f(t,x), \quad (' = \frac{d}{dt}) \tag{1.1.1}$$

with the initial condition

$$x(0) = \alpha . \tag{1.1.2}$$

Let us assume that the function f(t, x) is defined and continuous in the region $B = \{(t, x) \mid 0 \leq t \leq b, -\infty < x < \infty\}$: let us also assume that f (t, x) satisfies a Lipschitz condition in x, that is, that there exists a constant M such that the inequality

$$|f(t,x) - f(t,x^*)| \leq M |x - x^*| \tag{1.1.3}$$

*Numbers in the margin indicate pagination in the foreign text.

holds for any two pairs (t, x) and (t, x*) in B.

We know from the theory of differential equations that the initial-value problem (1.1.1, 2) has exactly one solution for $0 \leq t \leq b$. In what follows, this solution will be called the exact solution and denoted by x(t). It is continuous and differentiable in the interval mentioned.

The exact solution of (1.1.1, 2) will now be approximated at discrete points $t_n = nh$ (where n is a nonnegative integer, h is a positive number, and $t_n \in [0, b]$) with solutions of special difference equations. To this end, we shall use the class of linear difference equations proposed by Dahlquist [10] *

$$\sum_{\nu=0}^{k} \alpha_\nu \cdot x_{n+\nu} = h \cdot \sum_{\nu=0}^{k} \beta_\nu \cdot f_{n+\nu}, \quad (k \geq 2, \; n = 0, 1, \ldots), \tag{1.1.4}$$

where $f_m = f(t_m, x_m)$. The coefficients α_ν and β_ν (for $\nu = 0, 1, \ldots, k$) are real and independent of n with $\alpha_k \neq 0$. The positive parameter h is the difference step used in the procedure. Equation (1.1.4) defines the general linear k-step procedure of approximation of the solution of an initial-value problem of the form (1.1.1, 2). Such an approximation at a point $t = t_m$ will be denoted by x_m (as opposed to the exact solution at that point $x(t_m)$). /8

Since the present work is based to a considerable degree on linear difference equations, we shall present some of the more important properties of the solutions of such equations in the next section. These results are taken primarily from [16] and [20].

1.2. Linear difference equations.

Let k denote a natural number and let I denote a nonempty set of consecutive integers. Suppose that functions $F_n(z_0, z_1, \ldots, z_k)$ are defined for all $n \in I$ and all real values of the k + 1 variables $z_0, z_1, \ldots, z_k$. Then, an equation of the form

$$F_n(y_n, y_{n+1}, \ldots, y_{n+k}) = 0$$

is a difference equation of order k.

---

*Numbers in square brackets refer to the references at the end of the report.

Suppose that F_n is, for all $n \in I$, a linear homogeneous function of z_0, z_1, ..., z_k. Then, the corresponding difference equation is a linear homogeneous difference equation of order k.

Consider a linear nonhomogeneous difference equation of order k:

$$\sum_{\nu=0}^{k} \alpha_{\nu, n} y_{n+\nu} = c_{n+k}. \tag{1.2.1}$$

The coefficients $\alpha_{\nu, n}$ (for $\nu = 0, 1, \ldots, k$) and the quantities c_{n+k} are defined for all $n \in I$. In analogy with the theory of linear differential equations, the solution of the homogeneous equation

$$\sum_{\nu=0}^{k} \alpha_{\nu, n} y_{n+\nu} = 0 \tag{1.2.2}$$

is a certain linear combination of particular solutions. The general solution of the nonhomogeneous equation is then the sum of the general solution of the homogeneous equation and a particular solution of the nonhomogeneous equation.

Definition: Suppose that the sequences $\left\{y_n^{(\nu)}\right\}$ ($\nu = 1, \ldots, k$) are solutions of (1.2.2). Suppose that the relationship /9

$$\sum_{\nu=1}^{k} A_\nu y_n^{(\nu)} = 0$$

holds for $n - n_0 = 0, 1, \ldots, k - 1$ only when $(A_1, \ldots, A_k) = (0, \ldots, 0)$. Under this condition, these k solutions are said to be linearly independent at the point $n = n_0$. The set of solutions $\left\{y_n^{(\nu)}\right\}$ is called a fundamental system of solutions of the difference equation (1.2.2) at the point $n = n_0$.

The following two theorems, which we present without proof, provide information regarding the properties of the general solution of (1.2.2).

Theorem 1.1. Suppose that the sequences $\left\{y_n^{(\nu)}\right\}$ ($\nu = 1, \ldots, k$) constitute a fundamental system of (1.2.2) for $n = n_0$. Then this system remains fundamental for $n \geq n_0$ as long as $\alpha_{0, n} \neq 0$ for all $n \geq n_0$.

Theorem 1.2. Suppose that the sequences $\left\{ y_n^{(\nu)} \right\}$ ($\nu = 1, \ldots, k$) constitute a fundamental system of (1.2.2) at the point $n = n_0$. Then the general solution of this equation for $n \geq n_0$ has the form

$$y_n = \sum_{\nu=1}^{k} A_\nu \, y_n^{(\nu)} ,$$

as long as $\alpha_{0,n} \neq 0$. The quantities $A_1, \ldots, A_k$ denote suitable constants.

In the case of a linear difference equation with constant coefficients (that is, with coefficients independent of n), it is possible to exhibit explicitly a fundamental system at the point $n = n_0 = 0$. In place of (1.2.2), we write

$$\sum_{\nu=0}^{k} \alpha_\nu \, y_{n+\nu} = 0 \qquad (n = 0, 1, \ldots) \qquad (1.2.3)$$

and introduce (just as with a linear differential equation with constant coefficients) the corresponding characteristic polynomial

$$\rho(z) = \sum_{\nu=0}^{k} \alpha_\nu z^\nu$$

If z is a zero of $\rho(z)$, then the sequence $\{z^n\}$ is a solution of (1.2.3). For the general solution, we have the following theorem: /10

Theorem 1.3. Suppose that the characteristic polynomial $\rho(z)$ of (1.2.3) has distinct zeroes $z_1, \ldots, z_m$ (for $m \leq k$). Let $p_1, \ldots, p_m$ denote the corresponding multiplicities (with $\sum_{\nu=1}^{m} p_\nu = k$). Then, the sequences

$$\left\{z^n\right\} , \quad \left\{n . z^n\right\} , \quad \left\{n(n-1) . z^n\right\} , \ldots ,$$

$$\left\{n(n-1) \ldots (n-p_\nu+2) . z^n\right\} \qquad (\nu = 1, \ldots, m)$$

constitute a fundamental system of solutions of equation (1.2.3) at the point $n = 0$.

Corollary. If $\rho(z)$ has k simple zeros, then the general solution of (1.2.3) has the form

$$y_n = \sum_{\nu=1}^{k} A_\nu \cdot z_\nu^n , \qquad (1.2.4)$$

where the quantities $A_1, \ldots, A_k$ denote suitable constants (these are determined by the k "initial values" $y_0, \ldots, y_{k-1}$).

The zeros of $\rho(z)$ are in general complex numbers. In the case in which $\rho(z)$ has only real coefficients, if $z = r \exp i\varphi$ is a zero of $\varphi(z)$, so is $\bar{z} = r \exp(-i\varphi)$. This means that sequences whose elements are

$$r^n \exp in\varphi \quad \text{and} \quad r^n \exp(-in\varphi)$$

are solutions of (1.2.3). But since every linear combination of two solutions of this equation is also a solution, the sequences whose elements are

$$1/2\,(z^n + z^{-n}) = r^n \cos n\varphi$$

$$1/2i\,(z^n - z^{-n}) = r^n \sin n\varphi$$

are real solutions of (1.2.3). Therefore, if the coefficients in the equation are real, it is always possible to exhibit a real fundamental system.

With an eye to later application, let us again look at the difference equation (1.1.4), which we write in the form /11

$$x_{n+k} = h \cdot \frac{\beta_k}{\alpha_k} f_{n+k} + \frac{1}{\alpha_k}\left(h \sum_{\nu=0}^{k-1} \beta_\nu \cdot f_{n+\nu} - \sum_{\nu=0}^{k-1} \alpha_\nu x_{n+\nu}\right) = F(x_{n+k}) \qquad (1.2.5)$$

We are assuming that $\alpha_k \neq 0$. If $\beta_k = 0$, equation (1.2.5) represents a trivial equation for x_{n+k}. If $\beta_k \neq 0$, it is not in general possible to solve (1.2.5) for x_{n+k}, which needs to be found by an iterative procedure, beginning with a suitable initial approximation $x_{n+k}^{(0)}$:

$$x_{n+k}^{(m)} = F(x_{n+k}^{(m-1)}), \quad m = 1, \ldots .$$

By virtue of the assumption on f(t, x), the function F satisfies the inequality

$$|F(x_{n+k}) - F(x_{n+k}^{*})| \leq h \cdot M \cdot \left|\frac{\beta_k}{\alpha_k}\right| \cdot |x_{n+k} - x_{n+k}^{*}|$$

for arbitrary x_{n+k} and x^*_{n+k}. We know from the theory of iteration that the sequence $\left\{x_{n+k}^{(m)}\right\}$ converges to a unique limit if

$$h < \left| \frac{\alpha_k}{\beta_k} \right| \cdot M^{-1}.$$

The right-hand member of this inequality will be denoted by H in what follows. It is assumed that the setp h is always contained in the interval (O, H).

1.3. Dahlquist's theory.

We now continue the discussion of the linear difference procedure, introduced in section 1.1, for numerical integration of a given initial-value problem of the form (1.1.1, 2).

To measure the quality of the k-step procedure defined by (1.1.4), we assign to this difference equation the difference operator

$$L[x(t); h] = \sum_{\nu=0}^{k} (\alpha_\nu x(t+\nu h) - h\beta_\nu x'(t+\nu h)) \qquad (1.3.1)$$

For analytic functions, we may set / 12

$$L[x(t); h] = \sum_{\nu=0}^{\infty} C_\nu x^{(\nu)}(t) \cdot h^\nu , \quad x^{(0)}(t) = x(t), \qquad (1.3.2)$$

Then, we have

$$C_0 = \sum_{\nu=0}^{k} \alpha_\nu$$

$$C_1 = \sum_{\nu=0}^{k} (\nu \cdot \alpha_\nu - \beta_\nu) \qquad (1.3.3)$$

$$C_m = \frac{1}{m!} \sum_{\nu=1}^{k} \nu^m \cdot \alpha_\nu - \frac{1}{(m-1)!} \sum_{\nu=1}^{k} \nu^{m-1} \cdot \beta_\nu$$

$$(m = 2, 3, \ldots) \qquad (1.3.4)$$

Definition. The difference operator (1.3.1) is said to be of order p if in (1.3.2)

$$C_0 = C_1 = \ldots = C_p = 0, \text{ but } C_{p+1} \neq 0 \text{ int.}$$

To characterize the order, let us use the notation

$$L\,[z(t);\ h] = C(h^{p+1}).$$

The difference procedure (1.1.4) is completely determined by giving its two so-called characteristic polynomials $\rho(z)$ and $\sigma(z)$:

$$\rho(z) = \sum_{\nu=0}^{k} \alpha_\nu z^\nu \tag{1.3.5}$$

$$\sigma(z) = \sum_{\nu=0}^{k} \beta_\nu z^\nu \tag{1.3.6}$$

These two polynomials will serve as the basis for the discussion of the difference procedure of the corresponding difference operator. In what follows, we shall write briefly "the (ρ, σ) procedure" for the k-step procedure defined by the polynomials $\rho(z)$ and $\sigma(z)$.

Definition 1. Let $s_0, \ldots, s_{k-1}$ denote functions of h that are defined for all $h \in (0, H)$ (see section 1.2). Suppose that there exists a positive finite constant S such that /13

$$|s_\nu(h)| \leq S, \quad \nu = 0, \ldots k-1, \quad h \in (0,H).$$

Denote by x_n the solution of (1.1.4) defined by the initial values $s_0, \ldots, s_{k-1}$. Suppose that there exists, for all functions f(t, x) that satisfy the conditions of the existence theorem, a constant $S^* = S^*(S)$ that satisfies the inequality

$$\max_{n=0,1,\ldots} |x_n| \leq S^*, \qquad t_n \in [0,b],$$

for all $h \in (0, H)$. Then, the (ρ, σ) procedure is said to be stable (in the strong sense).

In different wording, this definition says that a difference procedure will be called stable if a solution that is defined by arbitrary bounded initial values and a sufficiently small step h remains uniformly bounded.

Definition 2. If x(t) is a solution of (1.1.1) and if, as $h \to 0$,

$$\operatorname{Max} \left| \sum_{\nu=0}^{k} (\alpha_\nu \; x(t_{n+\nu}) - h\beta_\nu \left(f(t_{n+\nu}, x(t_{n+\nu})) \right) \right| = O(h),$$

the (ρ, σ) procedure is said to be consistent.

Dahlquist has shown [5] that stability and consistency can be described by simple algebraic conditions on the characteristic polynomials.

Theorem 1.4. The (ρ,σ) procedure is stable if and only if all the zeros of $\rho(z)$ lie in or on the unit disk and those on the unit disk are simple.

Theorem 1.5. The (ρ, σ) procedure is consistent if and only if the order of the assigned difference operator is at least equal to 1. An equivalent condition is given by

$$\rho(1) = 0, \qquad \rho'(1) = \sigma(1).$$

Finally, let us confine the concept of convergence of the procedure (1.1.4): /14

Definition 3. Let $s_0, \ldots, s_{k-1}$ denote functions of h defined for all $h \in (0, H)$ and suppose that

$$\lim_{h \to 0} s_\nu(h) = x(0) = s \quad (\nu = 0, 1, \ldots, k-1).$$

Let x_n denote that solution of (1.1.4) that has the initial values $s_0, \ldots, s_{k-1}$. If

$$\operatorname{Max}_{n=0,1,\ldots} | x_n - x(t_n)| \to 0 \qquad \text{as} \quad h \to 0,$$

then the (ρ, σ) procedure is said to be convergent.

The question arises as to the conditions under which (ρ, σ) procedure is convergent. The answer to this question is also due to Dahlquist [10]:

Theorem 1.6. Stability and consistency are necessary and sufficient conditions for convergence of the (ρ, σ) procedure.

Dahlquist has also investigated the maximum order of a stable and consistent difference procedure:

Theorem 1.7. The order of a stable and consistent difference operator with given degree k cannot exceed $p = k + 2$. The maximum order is reached if and only if the characteristic polynomials $\rho(z)$ and $\sigma(z)$ satisfy the following conditions:

a) the degree k is an even number;

b) all zeros of $\rho(z)$ lie on the unit circle and are simple;

c) if

$$\frac{\rho(z)}{\log z} = \sum_{\nu=0}^{\infty} c_\nu \cdot (z-1)^\nu , \quad \log 1 = 0, \tag{1.3.8}$$

then $\sigma(z)$ is given by

$$\sigma(z) = \sum_{\nu=0}^{k} c_\nu \cdot (z-1)^\nu .$$

The following definition holds for the present work: /15

Definition 4. If the difference operator of a stable and consistent difference procedure of degree k is of maximum order $p = k + 2$, we shall call the operator an optimal difference operator and the procedure an optimal difference procedure.

For completeness, we shall give yet another theorem of Dahlquist regarding optimal difference procedures [10].

Theorem 1.8. If a difference procedure of even degree k is stable and consistent, the conditions

$$\alpha_\nu = -\alpha_{k-\nu} , \quad \beta_\nu = \beta_{k-\nu} \quad (\nu = 0,1, \ldots ,k) \tag{1.3.9}$$

are necessary and sufficient for optimality of the corresponding difference operator.

The best-known example of an optimal difference procedure is probably the Milne-Simpson procedure:

$$x_{n+2} - x_n = \frac{h}{3} (f_{n+2} + 4f_{n+1} + f_n),$$

with

$$\rho(z) = z^2 - 1, \quad \sigma(z) = \frac{1}{3} (z^2 + 4z + 1).$$

1.4. Marginal stability and growth parameters.

Marginal stability of an optimal difference procedure is described with the aid of the special initial-value problem

$$x' = -Ax, \quad x(0) = 1, \quad A > 0, \tag{1.4.1}$$

The corresponding difference equation is

$$\sum_{\nu=0}^{k} \alpha_\nu x_{n+\nu} = -Ah \sum_{\nu=0}^{k} \beta_\nu x_{n+\nu}. \tag{1.4.2}$$

This equation is a linear homogeneous equation with constant coefficients. For those solutions of the corresponding characteristic equation / 15

$$\varrho(z) + Ah \cdot \sigma(z) = 0,$$

that approximate the zeros z_ν of the polynomial $\rho(z)$, we set

$$\xi_\nu = z_\nu \cdot (1 - Ah \cdot \lambda_\nu + \mathcal{O}(h^2)), \quad \nu = 1, \ldots, j.$$

Thus, the (ρ, σ) procedure is taken to be optimal. We know from the theory of functions that the solutions are distinct for sufficiently small values of Ah. Thus, in accordance with Theorem 1.3, the general solution of (1.4.2) has, for sufficiently small values of Ah, the form

$$x_n = \sum_{\nu=1}^{k} C_\nu \cdot \exp(\ln \varphi_\nu) \cdot \exp(-\lambda_\nu A t_n) + \mathcal{O}(h) \tag{1.4.3}$$

$$(h \to 0, \quad nh = t_n).$$

Here, we have used the relationship

$$(1 - Ah \cdot \lambda_\nu + \mathcal{O}(h^2))^n = \exp(t_n \cdot h^{-1} (-\lambda_\nu Ah + \mathcal{O}(h^2)))$$

$$= \exp(-\lambda_\nu A t_n) + \mathcal{O}(h)$$

The undetermined coefficients λ_ν are given by

$$\lambda_\nu = \frac{\sigma(z_\nu)}{z_\nu \cdot \varrho'(z_\nu)} \tag{1.4.4}$$

Here, we have $\rho'(z_\nu) \neq 0$ (for $\nu = 1, \ldots, k$) since the (ρ, σ) procedure is assumed to be optimal and, by virtue of the consistency condition (1.3.7), $\lambda_1 = 1$ for $z_1 = 1$. Therefore, the first term in (1.4.3) approximates the exact solution of (1.4.1), which decreases exponentially as t increases. Since the (ρ, σ) procedure is optimal, we have $k \geq 2$ and $|z_\nu| = 1$ for $\nu = 1, \ldots, k$. Therefore, extraneous expressions appear in (1.4.3). As long as one of the quantities λ_ν is negative, these expressions increase exponentially and thus render the values of x_n meaningless. Dahlquist calls these λ_ν growth parameters and he has shown [11] that negative λ_ν appear with every optimal procedure. In other words, the solution of an optimal k-step procedure does not necessarily provide usable approximation values for the exact solution of a given initial-value problem. Optimal procedures are therefore called marginally stable.

We shall now give another important theorem regarding the growth parameters [11]:

Theorem 1.9. The growth parameters of an optimal difference procedure are real. For $\nu \neq 1$, the growth parameters have alternating signs:

$$\operatorname{sign}(\lambda_\nu) = (-1)^{\nu - \frac{k}{2}} \quad (\nu = 2, \ldots, k).$$

The growth parameter corresponding to the zero $z = -1$ satisfies the inequality $\lambda \leq -1/3$ with equality holding only for $k = 2$.

1.5. The purpose of the present work.

The purpose of this dissertation consists in modifying the optimal difference procedures proposed by Dahlquist in such a way that they no longer indicate marginal stability but still possess the maximum order.

In the next two chapters, modified characteristics of polynomials will be constructed whose coefficients are functions of the step h and a suitably chosen "stabilization parameter". It will be shown that the corresponding difference procedure converges and that it is an optimal procedure.

Chapter 3 discusses the choice of stabilization parameter for various linear and nonlinear differential equations. The concepts of absolute and relative stability will be defined.

In Chapter 4, these results will be extended to the case of systems of linear and nonlinear differential equations.

Chapter 5 lies somewhat off the path taken in the first four chapters. /16
There, explicit difference procedures of maximal order will be investigated for marginal stability and appropriate modifications will be proposed.

The remaining three chapters contain some examples of modified (implicit and explicit) difference procedures of maximum order (for $k = 2, 4,$ and 6). Finally, some numerical results are given for illustration purposes.

I should like to take this opportunity to express my heartfelt thanks to Professor Peter Henrici for his patience and for numerous suggestions. His ideas and his valuable criticisms were decisive for the completion of this work.

CHAPTER 2

MODIFIED OPTIMAL DIFFERENCE PROCEDURES

2.1. The modified characteristic polynomials.

For what follows, we assume that the degree of the polynomial $\rho(z)$ is exactly k and that the (ρ, σ) procedure is optimal. The zeros of $\rho(z)$ are then simple and they have the form

$$z_\nu = \exp(i\varphi_\nu) \qquad (\nu = 1, \ldots, k, \quad \varphi_\nu \in [0, 2\pi)). \tag{2.1.1}$$

The numbering is always according to increase in the argument φ_ν. It follows from the consistency condition (1.3.7) that $z_1 = 1$. Also,

$$z_{\frac{k}{2}+1} = -1,$$

since the coefficients in $\rho(z)$ are assumed real and k is an even number. The remaining k - 2 zeros come in complex-conjugate pairs:

$$\bar{z}_\nu = z_{k+2-\nu}, \quad k \geq 4 \qquad (\nu = 2, \ldots, \tfrac{k}{2}).$$

Let us now look at the conformal mapping

$$w = \frac{(2 - hL)z + hL}{2} \tag{2.1.2}$$

with $hL \in [0, 2)$. Here, h denotes the integration step, and L is a positive parameter, known as the stabilization parameter. This mapping maps the unit circle $|z| = 1$ onto the circle K*:

$$\left|w - \frac{hL}{2}\right| = 1 - \frac{hL}{2}$$

and it maps the point $z = 1$ into the point $w = 1$. Apart from this point, K^* is, for $hL \in (0, 2)$, completely in the interior of the circle $|w| = 1$.

The mapping inverse to (2.1.2) is given by

$$z = \frac{2w - hL}{2 - hL}, \qquad hL \notin [0,2), \tag{2.1.3}$$

The mapping (2.1.2) maps the zeros z_ν of $\rho(z)$ into the interior of the unit circle: /20

$$w_\nu = \frac{2 - hL}{2} \cdot z_\nu + \frac{hL}{2} = z_\nu + \frac{hL}{2}(1 - z_\nu) \quad (\nu = 1, \ldots, k). \tag{2.1.4}$$

Let us now take the numbers w_ν as the zeros of a new polynomial $\tilde{R}(w, hL)$. We have

$$\tilde{R}(w,hL) = \rho\left(\frac{2w - hL}{2 - hL}\right) = \sum_{\nu=0}^{k} \alpha_\nu \cdot \left(\frac{2w - hL}{2 - hL}\right)^\nu. \tag{2.1.5}$$

For $hL \in [0, 2)$, all the zeros of $\tilde{R}(w, hL)$ are of multiplicity 1. We could now take this polynomial as a new characteristic polynomial corresponding to the polynomial $\rho(z)$. The modified second characteristic polynomial $\tilde{S}(w, hL)$ could now be determined from it by application of the relation (1.3.8). However, for greater values of k, higher powers of hL could appear and, for this reason, the new procedure $(\tilde{R}, \tilde{S})$ might, because of its complicated structure, be of no practical significance. Therefore, let us take a different course. Instead of $\tilde{R}(w, hL)$, let us examine the polynomial $R(w, hL)$ linearized in hL. Since $hL/2 < 1$ and since

$$\frac{2w - hL}{2 - hL} = \left(w - \frac{hL}{2}\right) \sum_{\nu=0}^{\infty} \left(\frac{hL}{2}\right)^\nu$$

it follows from (2.1.5) that

$$\tilde{R}(w,hL) = \sum_{\nu=0}^{k} \alpha_\nu \left(\left(w - \frac{hL}{2}\right) \cdot \sum_{m=0}^{\infty} \left(\frac{hL}{2}\right)^m\right).$$

Linearization in hL then yields immediately

$$R(w,hL) = \sum_{\nu=0}^{k} \alpha_\nu \left(w^\nu + \nu \cdot \frac{hL}{2} \cdot (w-1) \cdot w^{\nu-1}\right)$$

or

$$R(\varpi, hL) = \varphi(\varpi) + \frac{hL}{2}\,\varphi^{\circ}(\varpi) = \sum_{\nu=0}^{k} \tilde{\alpha}_{\nu} \cdot \varpi^{\nu}\,, \tag{2.1.6}$$

where

$$\varphi^{\circ}(\varpi) = (\varpi - 1) \cdot \varphi'(\varpi) = \sum_{\nu=0}^{k} p_{\nu}\, \varpi^{\nu}\,. \tag{2.1.7}$$

Theorem 2.1. /21

(a) The polynomial $\rho^*(w)$ is of degree exactly k.

(b) All zeros of $\rho^*(w)$ lie inside the unit circle except for the zero at w = 1.

Proof:

(a) This is immediately obvious since

$$\varphi^{\circ}(\varpi) = k \cdot \alpha_k \varpi^k + \ldots - \alpha_1$$

and α_k is assumed to be nonzero.

(b) The second part of the theorem is an immediate consequence of the well-known theorem of Lucas ([14], p. 14):

Every convex polygon that contains all the zeros of a polynomial f(w) also contains all the zeros of its derivative f'(w). Since all the zeros of $\rho(z)$ are simple, $\rho'(w)$ can have no zeros on the unit circle. According to the theorem just cited, they lie in the interior of that circle.

Theorem 2.2. There exists a number $\Omega_1 > 0$ such that the zeros $w_\nu = w_\nu(hL)$ of the polynomial R(w, hL) are simple for all values of $hL \in [0, \Omega_1)$.

Proof: It follows from (2.1.6) that

$$R(1, hL) = 0 \quad (hL < 2). \tag{2.1.8}$$

The zero $w_1 = 1$ is independent of hL and, by virtue of the assumption regarding $\rho(z)$, it is simple.

To prove that the remaining zeros of R(w, hL) are simple for sufficiently small values of hL, let us use the following theorem from the theory of functions ([1], Theorem 11, p. 131):

Suppose that the function f(z) is analytic at a point $z = z_0$ and that $f(z_0) = w_0$. Suppose also that the function $f(z) - w_0$ has a zero of order n at z_0. Then, for sufficiently small $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$ such that the equation $f(z) = a$ has, for all a satisfying the inequality $|a - w_0| < \delta$, exactly n roots in the disk $|z - z_0| < \epsilon$.

In the present case, we take

$$f(z) = \frac{\varphi(z)}{(1 - z) \cdot \varphi'(z)} \cdot \qquad /22$$

By virtue of the assumption on $\rho(z)$, this function is analytic and the point $z = z_\nu$ (for $\nu = 2, \ldots, k$) and we have $f(z_\nu) = 0$. Every point z_ν is a zero of order 1. According to the theorem just stated, the equation $f(z) = hL/2$ has, for sufficiently small values of hL (such that $hL < \Omega_1$) exactly one root in the disk $|z - z_\nu| < \epsilon$ for $\nu = 2, \ldots, k$. This completes the proof.

Theorem 2.3. There exists a positive number Ω_2 such that the zeros $w_\nu = w_\nu(hL)$ (where $\nu = 2, \ldots, k$) of R(w, hL) lie, for all $hL \in (0, \Omega_2)$, in the interior of the unit circle.

Proof. The coefficients $\tilde{\alpha}_\nu$ of the polynomial R(w, hL) are, in accordance with (2.1.6), linear functions of hL:

$$\tilde{\alpha}_\nu = \alpha_\nu + \frac{hL}{2} \cdot p_\nu \qquad (\nu = 0, 1, \ldots, k).$$

The proof of this theorem (which is important for the convergence of the (R, S) procedure) rests to a considerable extent on the fact that the zeros $w_\nu(hL)$ of the polynomial R(w, hL) are continuous and continuously differentiable functions of hL, in particular, in a neighborhood of hL = 0. This implies the existence of

$$\left.\frac{dw_j}{d(hL)}\right|_{hL=0} = \dot{w}_j(0) \qquad (j = 2, \ldots, k).$$

Since

$$R(w_j, hL) = 0 = \sum_{\nu=0}^{k} \alpha_\nu w_j^\nu + \frac{hL}{2}(w_j - 1) \cdot \sum_{\nu=1}^{k} \nu \alpha_\nu w_j^{\nu-1}$$

we obtain after a simple calculation

$$\dot{w}_j(hL)\cdot\left[\sum_{\nu=1}^{k}\nu\,\alpha_\nu\,w_j^{\nu-1}(hL)+\frac{hL}{2}\left((w_j(hL)-1)\cdot\sum_{\nu=2}^{k}\nu(\nu-1)\alpha_\nu\,w_j^{\nu-2}(hL)\right.\right.$$
$$\left.\left.+\sum_{\nu=1}^{k}\nu\,\alpha_\nu\,w_j^{\nu-1}(hL)\right)\right]=-\frac{1}{2}\cdot(w_j(hL)-1)\cdot\sum_{\nu=1}^{k}\nu\alpha_\nu w_j^{\nu-1}(hL).$$

However, by assumption, $w_j(0)=z_j$ for $j=2,\ldots,k$ and, since all the zeros of $\rho(z)$ are simple, / 23

$$\sum_{\nu=1}^{k}\nu\cdot\alpha_\nu\cdot w_j^{\nu-1}(0)=\rho'(z_j)\neq 0.$$

It then follows that, obviously,

$$\dot{w}_j(0)=-\frac{1}{2}(z_j-1)\qquad(j=2,\ldots,k).$$

Let us take $w_j(hL)$ as a vector whose components are functions of the parameter hL. Then, as is shown in elementary differential geometry, the vector $w_j(0)$ represents a tangential vector of the curve described by $w_j(hL)$ at the point $w_j(0)=z_j$ and, in fact, it is a tangential vector in the direction of increase in the curve parameter hL. The proof will be complete if we show that this tangential vector points into the interior of the unit circle for $j=2,\ldots,k$. The components of $w_j(0)$ are

$$\frac{1}{2}(1-\cos\varphi_j)\text{ and }-\frac{1}{2}\sin\varphi_j.$$

Also, $w_j(0)\neq 0$ for all $j=2,\ldots,k$. The vector from z_j to z_1 has components $(1-\cos\varphi_j,\ -\sin\varphi_j)$, and a comparison of these two component pairs shows that the vector $w_j(0)$ does indeed point into the interior of the unit circle. But since $w_j(hL)$ is continuous in a neighborhood of $hL=0$, there exists an $\Omega_2>0$ such that the inequality $hL\in(0,\ \Omega_2)$ holds for $j=2,\ldots,k$ and all $|w_j(hL)|<1$. This completes the proof of the theorem.

The second modified characteristic polynomial S(w, hL) is constructed in accordance with (1.3.8). The function

$$\rho^*(w)\cdot(\log w)^{-1}\qquad(\text{with }\log 1=0)$$

is holomorphic at the point $w=1$ since $\rho^*(w)$ has, in accordance with Theorem 2.1, a simple zero there. Therefore, we again take

$$\rho^{*}(w) \cdot (\log w)^{-1} = \sum_{\nu=0}^{\infty} \gamma_\nu \cdot (w-1)^\nu \qquad (2.1.9)$$

and define

/24

$$\sigma^{*}(w) = \sum_{\nu=0}^{k} \gamma_\nu (w-1)^\nu = \sum_{\nu=0}^{k} q_\nu \cdot w^\nu . \qquad (2.1.10)$$

Then, let us define the modified polynomial S(w, hL) by

$$S(w, hL) = \sigma(w) + \frac{hL}{2} \cdot \sigma^{*}(w) = \sum_{\nu=0}^{k} \tilde{\beta}_\nu w^\nu . \qquad (2.1.11)$$

Equation (2.1.9) and (2.1.7) imply

$$\gamma_0 = \rho'(1) . \qquad (2.1.12)$$

For numerical integration of the initial-value problem (1.1.1.2), we now apply not the (ρ,σ) difference procedure but the modified (R, S) procedure:

$$\sum_{\nu=0}^{k} \tilde{\alpha}_\nu x_{n+\nu} = h \cdot \sum_{\nu=0}^{k} \tilde{\beta}_\nu \cdot f_{n+\nu} . \qquad (2.1.13)$$

The coefficients of the two assigned characteristic polynomials are linear functions of hL:

$$\tilde{\alpha}_\nu = \alpha_\nu + \frac{hL}{2} \cdot p_\nu$$

$$\tilde{\beta}_\nu = \beta_\nu + \frac{hL}{2} \cdot q_\nu \qquad (\nu = 0, 1, \ldots, k). \qquad (2.1.14)$$

In analogy with what was done in Chapter 1, let us assign to the (R, S) procedure a difference operator M. We set

$$M[x(t); hL] = L[x(t); h] + \frac{hL}{2} \cdot M^{*}[x(t); h] , \qquad (2.1.15)$$

where the new operator

$$M^{*}[x(t); h] = \sum_{\nu=0}^{k} \left(p_\nu \cdot x(t+\nu h) - h \cdot q_\nu \cdot x'(t+\nu h)\right) \qquad (2.1.16)$$

can be taken as the difference operator assigned to the (ρ^*, σ^*) procedure.

Theorem 2.4. The characteristic polynomials of the (R, S) difference procedure satisfy the relationships /25

$$R(1,hL) = 0 \qquad \text{and} \qquad R'(1,hL) = S(1,hL)$$
$$\text{for } hL \in [0,2). \tag{2.1.17}$$

In what follows, we shall call these relationships the consistency conditions of the (R, S) procedure.

Proof: The first relationship follows immediately from (2.1.8). For the derivative of R (w, hL) with respect to w, we find

$$R'(w,hL) = \varrho'(w) + \frac{hL}{2} \cdot (\varrho'(w) + (w-1) \cdot \varrho''(w)).$$

Thus, we have

$$R'(1,hL) = \varrho'(1).(1 + \frac{hL}{2}).$$

On the other side, by virtue of (2.1.11) and (2.1.12),

$$S(1,hL) = \sigma'(1) + \frac{hL}{2} \cdot \varrho'(1).$$

By assumption, (1.3.7) (the consistency condition for the (ρ, σ) procedure) holds, and this implies the equality of the right sides of the last two relationships.

2.2. The order of the difference operator M.

Define the function $\varphi^*(w)$ by

$$\varphi^\circ(w) = \frac{\varrho^\circ(w)}{\log w} - \sigma^\circ(w) \qquad (\log 1 = 0). \tag{2.2.1}$$

This function is holomorphic at the point w = 1 since $\rho^*(w)$ has a simple zero there. The polynomial $\sigma^*(w)$ was defined in such a way that the function $\varphi^*(w)$ has a zero of order p = k + 1 at the point w = 0. The following lemma, the proof of which can be found in [20] (p. 226), provides information regarding the order of the difference operator M*:

Lemma 2.1. Suppose that the above-described function $\varphi^*(w)$ has a zero of order exactly p = k + 1 at the point w = 1. Then, the difference operator M* defined by the polynomials $\rho^*(w)$ and $\sigma^*(w)$ is of order exactly p = k + 1. The converse of this assertion is also true. / 26

From this it follows, by virtue of the definition of the order of a difference operator that whenever

$$M^\circ[x(t);h] = \sum_{\nu=0}^{\infty} \cdot C_\nu^\circ \cdot h^\nu \cdot x^{(\nu)}(t)$$

we have

$$C_0^\circ = \ldots = C_{k+1}^\circ = 0, \quad C_{k+2}^\circ \neq 0$$

then,

$$M^\circ [x(t); h] = C_{k+2}^\circ \cdot x^{(k+2)}(t) \cdot h^{k+2} + O(h^{k+3}) \qquad (2.2.2)$$

for functions $x(t)$ that are at least $k + 3$ times continuously differentiable. With (1.3.4), we have

$$C_{k+2}^\circ = \frac{1}{(k+2)!} \sum_{\nu=1}^{k} \nu^{k+2} \cdot p_\nu - \frac{1}{(k+1)!} \sum_{\nu=1}^{k} \nu^{k+1} \cdot q_\nu . \qquad (2.2.3)$$

Since the given difference operator L is assumed to be optimal, the modified difference operator M has, for all sufficiently-many times continuously differentiable functions x (t), the form

$$M [x(t); hL] = (C_{k+3} \cdot x^{(k+3)}(t) + \frac{L}{2} \cdot C_{k+2}^\circ \cdot x^{(k+2)}(t)) \cdot h^{k+3} + O(h^{k+4}), \qquad (2.2.4)$$

where

$$C_{k+3} = \frac{1}{(k+3)!} \sum_{\nu=1}^{k} \nu^{k+3} \cdot \alpha_\nu - \frac{1}{(k+2)!} \sum_{\nu=1}^{k} \nu^{k+2} \cdot \beta_\nu \neq 0. \qquad (2.2.5)$$

We have thus proven

Theorem 2.5. The modified difference operator M is optimal: that is, its order, like that of the operator L, is the maximum order $p = k + 2$.

2.3. Convergence of the (R, S) procedure. /27

Definition. Let the quantity d (L) be defined for fixed $L \geq 0$ by

$$d(L) = \operatorname{Max}_{g \in D(L)} g \qquad (d(0) = H)$$

where

$$D(L) = \left\{ g \mid 0 < g, g \leq G(g) \right\}$$

and

$$G(h) = \frac{|\tilde{\alpha}_k (hL)|}{|\beta_k (hL)|} \cdot \frac{1}{M} .$$

The number M again denotes a Lipschitz constant of the function f (t, x). The set D(L) if a nonempty set for every $L \geq 0$ since $\alpha_k \neq 0$ and therefore

$$\tilde{\alpha}_k = \alpha_k + \frac{hL}{2} \cdot p_k = \alpha_k \cdot (1 + k \cdot \frac{hL}{2}) \neq 0 .$$

For what follows, let us assume that (for a fixed $L \geq 0$) the step h is contained in the interval (0, d(L)). This means, in particular, that

$$h \cdot |\tilde{\alpha}_k^{-1}| \cdot |\tilde{\beta}_k| \cdot M < 1 .$$

Let us also define $\Omega = \text{Min}(\Omega_1, \Omega_2)$, where Ω_1 and Ω_2 are as defined in Theorems 2.2 and 2.3.

Definition. Suppose that $s_0, \ldots, s_{k-1}$ are functions of h defined for all h x (0, d(L)) and suppose that

$$\lim_{h \to 0} s_\nu(h) = x(0) = s_\nu \quad (\nu = 0, 1, \ldots, k-1) .$$

Let x_n denote that solution of the difference equation (2.1.13) defined by the initial values $x_\nu = s_\nu(h)$ (for $\nu = 0, 1, \ldots, k-1$). Then, if

$$\max_{\substack{n=0,1,\ldots \\ t_n \in [0,b]}} |x_n - x(t_n)| \longrightarrow 0 , \qquad h \to 0,$$

the (R, S) procedure is said to be convergent. (We still denote the exact solution of (1.1.1.2) by x(t). /28

Theorem 2.6. The modified (R, S) procedure obtained from the optimal (ρ, σ) procedure is convergent.

Proof: Let the sequence $\{x_n\} = \{x_n(h)\}$ denote the solution equation (2.1.13) defined by the initial values $x_m = s_m(h)$ (for $m = 0, 1, \ldots, k-1$). If we set

$$\varepsilon(h) = \max_{m=0,\ldots,k-1} |s_m(h) - x(t_m)| = \max_{m=0,\ldots,k-1} |x_m(h) - x(t_m)| , \qquad (2.3.1)$$

then

$$\lim_{h \to 0} \varepsilon(h) = 0.$$

Let us now estimate the value of the expression M [x(t); hL]. By virtue of our assumption, the exact solution of (1.1.1.2) is continuous and differentiable on

[0, b], so that, by virtue of the theorem of the mean

$$x(t_{m+\nu}) = x(t_m) + \nu h \cdot x'(\xi_\nu),$$
$$t_m < \xi_\nu < t_{m+\nu}, \quad t_m \in [0,b]$$
$$t_{m+\nu} \in [0,b]. \qquad (2.3.2)$$

Furthermore,

$$x'(t_{m+\nu}) = x'(t_m) + \theta_m \operatorname{Max}_{|t-t^\circ| \leq \nu h} |x'(t) - x'(t^\circ)|$$
$$t, t^\circ \in [0,b] \quad (|\theta_m| \leq 1). \qquad (2.3.3)$$

If we not set

$$\operatorname{Max}_{\substack{|t-t^\circ| \leq \nu h \\ t, t^\circ \in [0,b]}} |x'(t) - x'(t^\circ)| = \chi(\nu h),$$

then the last two conditions become

$$x(t_{m+\nu}) = x(t_m) + \nu h \cdot \left[x'(t_m) + \theta'_m \cdot \chi(\nu h)\right] \quad (|\theta'_m| \leq 1) \qquad (2.3.4)$$

/29

and

$$x'(t_{m+\nu}) = x'(t_m) + \theta_m \cdot \chi(\nu h) \quad (|\theta_m| \leq 1). \qquad (2.3.5)$$

Let us write the difference operator M in the form

$$M\left[x(t_m); h\right] = \sum_{\nu=0}^{k} \alpha_\nu \, x(t_m + \nu h) - h \sum_{\nu=0}^{k} \beta_\nu \, x'(t_m + \nu h)$$
$$+ \frac{hL}{2} \left(\sum_{\nu=0}^{k} p_\nu \, x(t_m + \nu h) - h \sum_{\nu=0}^{k} q_\nu \, x'(t_m + \nu h) \right).$$

With the aid of the relationships (2.3.4) and (2.3.5), we find

$$M\left[x(t_m); hL\right] = x(t_m) \cdot \sum_{\nu=0}^{k} \alpha_\nu + h \cdot x'(t_m) \cdot \sum_{\nu=0}^{k} (\nu \alpha_\nu - \beta_\nu)$$
$$+ h \cdot \theta'_m \cdot \sum_{\nu=0}^{k} \nu \alpha_\nu \chi(\nu h) - h \cdot \theta_m \cdot \sum_{\nu=0}^{k} \beta_\nu \chi(\nu h) \qquad (2.3.6)$$
$$+ \frac{hL}{2} \left[x(t_m) \cdot \sum_{\nu=0}^{k} p_\nu + h \, x'(t_m) \cdot \sum_{\nu=0}^{k} (\nu p_\nu - q_\nu) \right.$$

$$+ h \cdot \theta'_m \sum_{\nu=0}^{k} \nu p_\nu \chi(\nu h) - h\theta_m \cdot \sum_{\nu=0}^{k} q_\nu \chi(\nu h) \Big] . \qquad (2.3.6)$$
(cont'd)

If we observe the consistency conditions (2.1.17), it then follows that

$$M[x(t_m); hL] = h\,(\theta'_m \cdot \sum_{\nu=0}^{k} \nu \alpha_\nu \cdot \chi(\nu h) - \theta_m \cdot \sum_{\nu=0}^{k} \beta_\nu \chi(\nu h)$$
$$+ h^2 \cdot \frac{L}{2}\,(\theta'_m \cdot \sum_{\nu=0}^{k} \nu p_\nu \cdot \chi(\nu h) - \theta_m \cdot \sum_{\nu=0}^{k} q_\nu \chi(\nu h)\,). \qquad (2.3.7)$$

Since

/34

$$\chi(\nu h) = \max_{\substack{|t-t^*| \le \nu h \\ t,t^* \in [0,b]}} |x'(t) - x'(t^*)|$$

we have

$$\chi(\nu h) \ge 0, \quad \chi(0) = 0, \quad \chi(\nu h) \le \chi(kh) \quad (\nu = 0,\ldots,k),$$

and then, as a direct consequence,

$$\left|\sum_{\nu=0}^{k} \nu\alpha_\nu \chi(\nu h)\right| \le \chi(kh) \cdot \sum_{\nu=0}^{k} \nu \cdot |\alpha_\nu|$$
$$\left|\sum_{\nu=0}^{k} \beta_\nu \chi(\nu h)\right| \le \chi(kh) \cdot \sum_{\nu=0}^{k} |\beta_\nu|$$
$$\left|\sum_{\nu=0}^{k} \nu \cdot p_\nu \chi(\nu h)\right| \le \chi(kh) \cdot \sum_{\nu=0}^{k} \nu\,|p_\nu|$$
$$\left|\sum_{\nu=0}^{k} q_\nu \cdot \chi(\nu h)\right| \le \chi(kh) \cdot \sum_{\nu=0}^{k} |q_\nu| .$$

Furthermore,

$$\theta'_m \cdot \sum_{\nu=0}^{k} \nu\alpha_\nu \chi(\nu h) = \theta' \cdot \chi(kh) \cdot \sum_{\nu=0}^{k} \nu\,|\alpha_\nu|, (|\theta'| \le 1)$$
$$\theta_m \cdot \sum_{\nu=0}^{k} \beta_\nu \chi(\nu h) = \theta \cdot \chi(kh) \cdot \sum_{\nu=0}^{k} |\beta_\nu| \quad (|\theta| \le 1)$$
$$\theta'_m \cdot \sum_{\nu=0}^{k} \nu p_\nu \chi(\nu h) = \vartheta' \cdot \chi(kh) \cdot \sum_{\nu=0}^{k} \nu |p_\nu|, \; (|\vartheta'| \le 1)$$

$$\theta_m \cdot \sum_{\nu=0}^{k} q_\nu \, \chi(\nu h) \quad = \quad \vartheta \cdot \chi(kh) \cdot \sum_{\nu=0}^{k} |q_\nu| \, , \quad (|\vartheta| \leq 1).$$

If we substitute these expressions into (2.3.7), that relationship reduces to

$$\left| M \left[x(t_m); hL \right] \right| \quad = \quad - \quad T_m \cdot h \cdot \chi(kh) \cdot \tilde{K} \quad (|T_m| \leq 1) \qquad /31$$

with the abbreviations

$$K \quad = \quad \sum_{\nu=0}^{k} (\nu \cdot |\alpha_\nu| \quad + \quad |\beta_\nu| \,)$$

$$K^\circ \quad = \quad \sum_{\nu=0}^{k} (\nu \, |p_\nu| \quad + \quad |q_\nu|)$$

$$K \quad = \quad K + \frac{hL}{2} \cdot K^\circ \, .$$

The second part of the proof consists in finding a difference equation for the discretization error $e_n = x_n (t_n)$ and then estimating the solution of this equation. We subtract from the difference equation (2.1.13) the corresponding expression $M [x(t); hL]$.

We obtain

$$\sum_{\nu=0}^{k} \tilde{\alpha}_\nu \, e_{m+\nu} - h \sum_{\nu=0}^{k} \tilde{\beta}_\nu \, g_{m+\nu} \cdot e_{m+\nu} = T_m \cdot h \cdot \chi(kh) \cdot \tilde{K}, \qquad (2.3.8)$$

where the quantity $g_n = g (t_n)$ is defined by

$$g_n = \begin{cases} \left[f(t_n, x_n) - f(t_n, x(t_n)) \right] \cdot e_n^{-1} \, , & e_n \neq 0 \\ 0 \, , & e_n = 0 \end{cases}$$

To estimate the solution of the difference equation (2.3.8), let us define

$$Q(w, hL) = R(w^{-1}, hL) \cdot w^k = \sum_{\nu=0}^{k} \tilde{\alpha}_\nu \, w^{k-\nu} \, . \qquad (2.3.9)$$

The zeros of this polynomial are

$$w_\nu = w_\nu^{-1}$$

$$(\nu = 1, \ldots, k)$$

Here, the quantities $w_\nu = w_\nu (hL)$, for $\nu = 1, \ldots, k$, are the zero of the polynomial $R(r, hL)$. For $hL \in [0, \Omega)$ the rational function $[Q(w, hL)]^{-1}$ has only simple poles at /32

$w = W_\nu$ for $\nu = 1, \ldots, k$. The decomposition of this function into partial fractions is therefore

$$[Q(w, hL)]^{-1} = \sum_{\nu=1}^{k} \frac{B_\nu}{w - w_\nu^{-1}} \qquad (2.3.10)$$

Furthermore,

$$Q(w, hL) = \tilde{\alpha}_0 \cdot \prod_{\nu=1}^{k} (w - w_\nu^{-1}),$$

where the cofficient $\tilde{\alpha}_0$ is nonzero for sufficiently small values of hL since

$$\tilde{\alpha}_0 = \alpha_0 + \frac{hL}{2} \cdot p_0, \qquad p_0 = \alpha_1 - 2\alpha_2,$$

and, in accordance with Theorem 1.8,

$$\alpha_0 = -\alpha_k \neq 0.$$

The coefficients B_ν in (2.3.10) have the form

$$B_\nu = \left(\tilde{\alpha}_0 \prod_{\substack{\mu=1 \\ \mu \neq \nu}}^{k} (w_\nu^{-1} - w_\mu^{-1})\right)^{-1} \quad (\nu = 1, \ldots, k).$$

It follows that

/33

$$[Q(w, hL)]^{-1} = \tilde{\alpha}_0^{-1} \cdot \sum_{\nu=1}^{k} \frac{B_\nu^\circ}{w - w_\nu^{-1}} \qquad (2.3.11)$$

where

$$B_\nu^\circ = \tilde{\alpha}_0 \cdot B_\nu \quad (\nu = 1, \ldots, k).$$

Since the relationship

$$\frac{1}{w - w_\nu^{-1}} = -w_\nu \cdot \sum_{m=0}^{\infty} (w \cdot w_\nu)^m$$

may be used for $|w \cdot w_\nu| < 1$, let us write, in place of (2.3.11),

$$[Q(w, hL)]^{-1} = -\tilde{\alpha}_0^{-1} \cdot \sum_{\nu=1}^{k} B_\nu^\circ \cdot w_\nu \left(\sum_{m=0}^{\infty} (w \cdot w_\nu)^m\right).$$

If we now set

$$[Q(w,hL)]^{-1} = \sum_{\nu=0}^{\infty} \tilde{\gamma}_\nu \cdot w^\nu \,, \qquad (2.3.12)$$

then

$$\tilde{\gamma}_n = -\tilde{\alpha}_k^{-1} \cdot \sum_{\nu=1}^{k} \tilde{B}_\nu^0 \cdot \tilde{\gamma}_{n+\nu-1} \,.$$

For all $w \in (0, w_0)$, we have

$$\max_{m} |\tilde{\gamma}_m| \leq |\tilde{\alpha}_k^{-1}| \cdot \sum_{\nu=1}^{k} |\tilde{B}_\nu^0| \,. \qquad (2.3.13)$$

We now find from (2.3.9) and (2.3.12)

$$1 = (\tilde{\alpha}_k + \tilde{\alpha}_{k-1} \cdot w + \ldots + \tilde{\alpha}_0 \cdot w^k)(\tilde{\gamma}_0 + \tilde{\gamma}_1 \cdot w + \ldots),$$

This immediately yields the following relationship between the coefficients $\tilde{\alpha}_\nu$ and $\tilde{\gamma}_\nu$.

$$\begin{aligned} &\tilde{\alpha}_k \cdot \tilde{\gamma}_0 = 1 \,, \\ &\tilde{\alpha}_k \cdot \tilde{\gamma}_\nu + \ldots + \tilde{\alpha}_0 \cdot \tilde{\gamma}_{\nu-k} = 0 \,. \quad (\nu > 0) \\ &(\tilde{\gamma}_\mu = 0 \text{ für } \mu < 0) \,. \end{aligned} \qquad (2.3.14)$$

For what follows, let us write the difference equation (2.3.8) in the form

$$\sum_{\nu=0}^{k} \tilde{\alpha}_\nu \, o_{m+\nu} = h \sum_{\nu=0}^{k} \tilde{\beta}_\nu \, \sigma_{m+\nu} \, o_{m+\nu} + l_m \quad (l_m = T_m \cdot h \, \chi(hh) \cdot \tilde{K}).$$

Let us now multiply the mth, that is, the (n - k - r)th, equation by $\tilde{\gamma}_r$, for r = 0, ..., n - k, and sum the results. A simple transformation yields the following expression for the left-hand member:

$$\Lambda_n = e_n + (\tilde{\alpha}_{k-1} \tilde{\gamma}_{n-k} + \ldots + \tilde{\alpha}_0 \tilde{\gamma}_{n-2k+1}) \, e_{k-1} + \ldots + \tilde{\alpha}_0 \tilde{\gamma}_{n-k} \, e_0 \,. \qquad (2.3.15)$$ /34

After an analogous transformation, the right-hand member becomes

$$\begin{aligned} R_n = h\Big[&\tilde{\beta}_k \tilde{\gamma}_0 \sigma_n e_n + (\tilde{\beta}_k \tilde{\gamma}_1 + \tilde{\beta}_{k-1} \tilde{\gamma}_0) \sigma_{n-1} \cdot e_{n-1} + \ldots \\ &+ (\tilde{\beta}_k \tilde{\gamma}_k + \ldots + \tilde{\beta}_0 \tilde{\gamma}_0) \, \sigma_{n-k} \cdot e_{n-k} + \ldots + \tilde{\beta}_0 \tilde{\gamma}_{n-k} \sigma_0 e_0 \Big] \\ &+ l_{n-k} \tilde{\gamma}_0 + \ldots + l_0 \tilde{\gamma}_{n-k} \,. \end{aligned}$$

When we equate Λ_n and R_n, we get

$$e_n = -\left[(\tilde{\alpha}_{k-1}\tilde{\gamma}_{n-k} + \dots + \tilde{\alpha}_0\tilde{\gamma}_{n-2k+1})e_{k-1} + \dots + \tilde{\alpha}_0\tilde{\gamma}_{n-k}e_0\right] +$$
$$h\left[\tilde{\beta}_k\tilde{\gamma}_0\sigma_n e_n + \dots + \tilde{\beta}_0\tilde{\gamma}_{n-k}\sigma_0 e_0\right] +$$
$$l_{n-k}\tilde{\gamma}_0 + \dots + l_0\tilde{\gamma}_{n-k}.$$

By virtue of our earlier assumptions and definitions, we have

$$\sup |\tilde{\gamma}_m| = \tilde{\Gamma}$$

$$\max_{m=0,\dots,k-1} |e_m| = \sigma(h)$$

$$|\varepsilon_m| = \frac{|f(t_m,x_m) - f(t_m,x(t_m))|}{|x_m - x(t_m)|} \leq M.$$

We thus obtain the estimate

$$|e_n| \leq \tilde{\Gamma}\cdot\sigma\cdot k\sum_{\nu=0}^{k}|\tilde{\alpha}_\nu| + h\cdot M\cdot|\tilde{\beta}_k|\cdot|\tilde{\gamma}_0|\cdot|e_n| +$$
$$M\cdot\tilde{\Gamma}\cdot h\sum_{\nu=0}^{k}|\tilde{\beta}_\nu|\cdot\sum_{\nu=0}^{n-1}|e_\nu| + \tilde{\Gamma}\cdot\sum_{\nu=0}^{n-k}|l_\nu|.$$

But since $\tilde{\alpha}_k\tilde{\gamma}_0 = 1$ whenever (2.3.14) holds, we have /35

$$|e_n| \leq hM\cdot|\tilde{\alpha}_k^{-1}\cdot\tilde{\beta}_k|\cdot|e_n| + M\cdot\tilde{\Gamma}\cdot h\cdot\tilde{B}\cdot\sum_{\nu=0}^{n-1}|e_\nu| \qquad (2.3.16)$$
$$+ \tilde{\Gamma}\cdot\sigma\cdot k\cdot\tilde{A} + N\tilde{\Gamma}\Lambda$$

with

$$\tilde{A} = \sum_{\nu=0}^{k}|\tilde{\alpha}_\nu|,$$

$$\tilde{B} = \sum_{\nu=0}^{k}|\tilde{\beta}_\nu|,$$

$$\max_{n=0,\dots,N}|l_n| \leq h\cdot\chi(kh)\cdot\tilde{K} = \Lambda.$$

The discretization error therefore satisfies the inequality

$$|e_n| \leq \frac{\tilde{\Gamma}}{1-hM\cdot|\tilde{\alpha}_k^{-1}\cdot\tilde{\beta}_k|}\left(M\cdot h\tilde{B}\cdot\sum_{\nu=0}^{n-1}|e_\nu| + k\cdot\sigma\cdot\tilde{A} + N\cdot\Lambda\right). \qquad (2.3.17)$$

$$(1 - hM|\tilde{\alpha}_k^{-1}\cdot\tilde{\beta}_k| < 1).$$

To obtain a somewhat more transparent form, let us make the following abbreviations:

$$\Gamma^{\circ} = \tilde{\Gamma} \cdot (1 - hM \cdot |\tilde{\alpha}_k^{-1}| \cdot |\tilde{\beta}_k|)^{-1},$$

$$B_1 = \tilde{B}\Gamma^{\circ},$$

$$M^{\circ} = M \cdot B_1,$$

$$H^{\circ} = \Gamma^{\circ} \cdot (k \cdot \sigma \cdot \tilde{A} + N \cdot \Lambda).$$

Then, instead of (2.3.17), we have

$$|e_n| \leq hM^{\circ} \cdot \sum_{\nu=0}^{n-1} |e_\nu| + H^{\circ}. \qquad (2.3.18)$$

/36

The general validity of this inequality can be shown by induction on n. Since

$$\tilde{A} \cdot \Gamma^{\circ} \geq 1$$

and

$$H^{\circ} \geq \Gamma^{\circ} \tilde{A} \cdot k \cdot \sigma$$

it follows that

$$H^{\circ} \geq \sigma. \qquad (2.3.19)$$

By assumption, we have for m = 0, 1, ..., k - 1

$$|e_m| \leq hm\sigma M^{\circ} + H^{\circ}$$

or, by virtue of (2.3.19),

$$|e_m| \leq H^{\circ}(1 + hM^{\circ})^m. \qquad (2.3.20)$$

Let us assume that the relationship above holds for m = 0, 1, ..., n - 1. When we use (2.3.18), it follows that

$$|e_n| \leq h \cdot M^{\circ}H^{\circ} \frac{(1 + hM^{\circ})^n - 1}{1 + hM^{\circ} - 1} + H^{\circ}.$$

We than get directly

$$|e_n| \leq H^{\circ} \cdot (1 + hM^{\circ})^n.$$

When we apply the familiar inequality

$$1 + hM^{\circ} \leq \exp(hM^{\circ})$$

it follows finally that

$$|e_n| \leq \Gamma^{\circ} \cdot \exp(nhM^{\circ}) .$$

Written out, this inequality has the form /37

$$|e_n| \leq \Gamma^{\circ} \cdot (\tilde{A}(h) \cdot k \cdot s(h) + N \cdot h \cdot \chi(kh) \cdot \tilde{K}(h)) \cdot \exp(t_n M \tilde{B}(h) \cdot \Gamma^{\circ}(h)) . \qquad (2.3.21)$$

It still remains to discuss the behavior of e_n as $h \to 0$ with $nh = t_n$ and L fixed. Since $\tilde{A}$, $\tilde{B}$, and $\tilde{K}$ are linear functions of h, they remain bounded. Furthermore, by assumption.

$$\lim_{h \to 0} s(h) = 0 ,$$

By virtue of the uniform continunity of x' (t) (which = f (t, x, (t)) on [0, b], we have

$$\lim_{h \to 0} \chi(kh) = \lim_{h \to 0} \left(\max_{|t - t^{\circ}| \leq k \cdot h} |x'(t) - x'(t^{\circ})| \right) = 0 .$$

It only remains to investigate the behavior of Γ° and of $\tilde{\Gamma}$. By virtue of (2.3.13), we have

$$\tilde{\Gamma} \leq \left|\tilde{\alpha}_0^{-1}\right| \cdot \sum_{\mu=1}^{k} \left(\prod_{\substack{\nu=1 \\ \nu \neq \mu}}^{k} \frac{|w_\nu \cdot w_\mu|}{|w_\nu - w_\mu|} \right) .$$

But, for sufficiently small values of hL, we have, on the one hand,

$$w_\nu \neq w_\mu \qquad (\nu \neq \mu).$$

and, on the other,

$$\tilde{\alpha}_0 \neq 0$$

Therefore, $\tilde{\Gamma}$ is bounded:

$$\lim_{h \to 0} \tilde{\Gamma}(h) = c < \infty$$

/38

Therefore,

$$|e_n| \to 0 \quad \text{as} \quad h \to 0 , \quad \text{as} = t_n \text{ fixed, L fixed } t_n \in [0, b] ,$$

which completes the proof of the theorem.

Another assertion regarding the discretization error, which we shall we use later, can be inferred from the convergence proof without further difficulty.

Let us assume that the values x_n do not satisfy the difference equation (2.1.13) exactly. Suppose that instead

$$\sum_{\nu=0}^{k} \tilde{\alpha}_\nu x_{n+\nu} - h\sum_{\nu=0}^{k} \tilde{\beta}_\nu f_{n+\nu} = \vartheta_n \cdot K_1 \cdot h^{q+1}$$

where $|\vartheta_n| < 1$, $K_1 \geq 0$, and $q > 0$. We then find the following estimate for the discretization error:

$$|e_n| \leq \Gamma^* \cdot (\tilde{A} \cdot k \cdot \varepsilon + N(h \cdot \chi(h) \cdot \tilde{K} + K_1 \cdot h^q)) \cdot \exp(t_n M \tilde{B} \Gamma^*). \qquad (2.3.22)$$

2.4. The constant of the (R, S) procedure.

The quantity

$$C = \frac{C_{k+3}}{\rho'(1)} = \frac{C_{k+3}}{\sum_{\nu=0}^{k} \beta_\nu},$$

which is invariant under multiplication of the difference equation (1.1.4) by a zero constant, is called the error constant of the (ρ, σ) procedure. Since $z_1 = 1$ is a simple zero of $\rho(z)$, we have $\rho'(1) \neq 0$.

Another expression for C can be found by applying the difference operator L to the function $x(t) = \exp t$. Since L is of order $p = k + 2$, we have

$$L[e^t; h] = e^t \cdot (\rho(e^h) - h \cdot \sigma(e^h)) = e^t \cdot C_{k+3} \cdot h^{k+3} + O(h^{k+4})$$

and hence

$$\rho(e^h) - h \cdot \sigma(e^h) = C_{k+3} \cdot h^{k+3} + O(h^{k+4}).$$

On the other hand, since $\rho \exp h - h \cdot \sigma \exp h$ has, by the construction of $\sigma(z)$, a zero of order $k + 3$ at the point $h = 0$ and is holomorphic there, we have

$$\frac{\rho(z)}{\log z} - \sigma(z) = C_{k+3} \cdot (z-1)^{k+2} + \sigma((z-1)^{k+3})$$

$$(h = \log z,\ \log z = (z-1) + O((z-1)^2)).$$

From this, we can conclude that the coefficient C_{k+3} in the expansion (1.3.2) and the first nonzero coefficient in (1.3.8) are equal (see [20], p. 227):

$$C = \frac{C_{k+3}}{\rho'(1)} = \frac{c_{k+2}}{\rho'(1)}$$

$$\sigma(1) = \rho'(1) = c_0,$$

or, since $\sigma(1) = \rho'(1) = c_0$,

$$C = \frac{c_{k+2}}{c_0} . \qquad (2.4.1)$$

In an analogous manner, let us now define the error constant C^* assigned to the (ρ^*, σ^*) procedure. When (2.2.2) holds, we have

$$M^\circ [e^t; h] = e^t \cdot (\rho^\circ(e^h) - h \cdot \sigma^\circ(e^h)) = C^\circ_{k+2} \cdot e^t \cdot h^{k+2} + O(h^{k+3}) .$$

Therefore, the function

$$\rho^*(\exp h) - h \cdot \sigma^*(\exp h)$$

has a zero of order exactly $k + 2$ at $h = 0$. Hence, with $z = \exp h$ and $\log z = z - 1 + O(z - 1^2)$, we have

$$\frac{\rho^\circ(z)}{\log z} - \sigma^\circ(z) = C^\circ_{k+2} \cdot (z - 1)^{k+1} + O((z - 1)^{k+2}).$$

Since, by virtue of (2.1.12), $\rho'(1) = \gamma_0$, let us define

$$(\frac{L}{2})^{-1} \cdot C^\circ = \frac{C^\circ_{k+2}}{\rho'(1)} = \frac{\gamma_{k+1}}{\gamma_0}$$

/ 40

or, in analogy with (2.4.1),

$$C^\circ = \frac{L}{2} \cdot \frac{C^\circ_{k+2}}{\sum_{\nu=0}^{k} \beta_\nu} .$$

The expressions for C and C^* written out are

$$C = \left(\sum_{\nu=0}^{k} \beta_\nu \right)^{-1} \cdot \left[\frac{1}{(k+3)!} \sum_{\nu=1}^{k} \nu^{k+3} \cdot \alpha_\nu - \frac{1}{(k+2)!} \sum_{\nu=1}^{k} \nu^{k+2} \cdot \beta_\nu \right]$$

and

$$C^\circ = \frac{L}{2} \cdot \left(\sum_{\nu=0}^{k} \beta_\nu \right)^{-1} \cdot \left[\frac{1}{(k+2)!} \sum_{\nu=1}^{k} \nu^{k-2} \cdot p_\nu - \frac{1}{(k+1)!} \sum_{\nu=1}^{k} \nu^{k+1} \cdot q_\nu \right] .$$

CHAPTER 3

/41

STABILIZATION

3.1. Stabilization in the case $x' = -Ax$.

We shall now determine the growth parapeters $\tilde{\lambda}_\nu$ corresponding to the (R, S) difference procedure. To this end, we introduce the special initial-value problem

$$x' = -Ax \quad , \quad x(0) = 1 \, , \tag{3.1.1}$$

where A denotes an arbitray nonzero complex constant. The difference equation (2.1.13) then takes the form

$$\sum_{\nu=0}^{k} \tilde{\alpha}_\nu \, x_{n+\nu} + Ah \sum_{\nu=0}^{k} \tilde{\beta}_\nu \, x_{n+\nu} = 0 \tag{3.1.2}$$

It is linear and homogeneous and it has constant coefficients. The solutions v_ν of the corresponding characteristic equation

$$R(w, hL) + Ah \cdot S(w, hL) = 0 \tag{3.1.3}$$

approximate the zeroes of the polynomial R (w, hL). Therefore, as in section 1.4, we set

$$v_\nu = w_\nu \, (1 - Ah \, \tilde{\lambda}_\nu + O(h^2)) \qquad (\nu = 1, \ldots, k) \tag{3.1.4}$$

When we substitute this expression into (3.1.3) and compare the coefficients of h, we obtain for the modified growth parameters

$$\tilde{\lambda}_\nu = \frac{S(w_\nu, hL)}{w_\nu \cdot R'(w_\nu, hL)} \qquad (\nu = 1, \ldots, k) \tag{3.1.5}$$

with $\tilde{\lambda}_1 = \lambda_1 = 1$.

For small values of hL, the $\tilde{\lambda}_\nu$ are bounded since, in accordance with Theorem 2.2, all the zeros w_ν are then simple and nonzero. Furthermore,

(under the assumption that $\rho(z)$ and $\sigma(z)$ have no common factors), $\tilde{\lambda}_\nu \neq 0$ for all $\nu = 1, \ldots, k$.

If we expand the expression (3.1.5) for $hL(0, \Omega)$ in powers of h, we see /42

that

$$\tilde{\lambda}_\nu = \lambda_\nu + \sigma(h) , \qquad (3.1.6)$$

where the quantities

$$\lambda_\nu = \frac{\sigma(z_\nu)}{z_\nu \cdot \rho'(z_\nu)} \quad (\nu = 1, \ldots, k)$$

are growth parameters of the (ρ, σ) procedure.

For the following discussion, let us refine somewhat the concept of stability. We shall follow the terminology of Hull and Newbery [27] and shall distinguish between relative and absolute stability:

Definition 1. The (R, S) difference procedure is said to be relatively stable with respect to (3.1.1) if the solutions of the characteristic equation (3.1.3) satisfy the condition

$$|v_\nu| < |v_1| \qquad (\nu = 2, \ldots, k)$$

Definition 2. The (R, S) difference procedure is said to be absolutely stable with respect to (3.1.1) if the solutions of the characteristic equation (3.1.3) satisfy the condition

$$|v_\nu| < 1 \qquad (\nu = 2, \ldots, k)$$

Our next task is to find conditions on the stabilization parameter L under which the (R, S) procedure with respect to the initial-value problem (3.1.1) is relatively or absolutely stable. Since

$$w_\nu = z_\nu \cdot \left(1 + \frac{hL}{2}\,(1 - z_\nu) \cdot z_\nu^{-1} + \sigma(h^2)\right)$$

we find that, instead of (3.1.4), we have

$$v_\nu^n = \exp(\ln \varphi_\nu) \cdot \left(1 + \frac{hL}{2}(1 - z_\nu) \cdot z_\nu^{-1} + \sigma(h^2)\right)^n \cdot (1 - Ah\,\tilde{\lambda}_\nu + \sigma(h^2))^n .$$

as $h \to 0$ with $nh = t_n$, L fixed. When we apply (3.1.6), we get immediately /43

$$v_\nu^n = \exp(\ln \varphi_\nu) \cdot \exp\left[\left(\frac{L}{2} \cdot \frac{1 - z_\nu}{z_\nu} - A\lambda_\nu\right) t_n\right] + \sigma(h) . \qquad (3.1.7)$$

Therefore, for relative stability, the condition

$$\exp\left(\frac{L}{2} \cdot \mathrm{Re}\,\frac{1 - z_\nu}{z_\nu} - \mathrm{Re}\,A\,\lambda_\nu\right) < \exp(-\mathrm{Re}A) \qquad (\nu = 2, \ldots, k)$$

must hold. This condition can be simplified since, in accordance with Theorem 1.9, the growth parameters of the optimal (ρ, σ) procedure are real. We find

$$\frac{L}{2} \cdot (\cos\varphi_\nu - 1) - \lambda_\nu\,\mathrm{Re}A < -\mathrm{Re}A \qquad (\nu = 2, \ldots, k).$$

In an analogous manner, we obtain the condition for absolute stability

$$\frac{L}{2} \cdot (\cos\varphi_\nu - 1) - \lambda_\nu\,\mathrm{Re}A, < 0 \qquad (\nu = 2, \ldots, k).$$

Let us express these results in the following theroems:

Theorem 3.1.a: If the stabilization parameter L satisfies the inequality

$$L > 2 \cdot \mathrm{Max}\,\frac{-(\lambda_\nu - 1) \cdot \mathrm{Re}A}{1 - \cos\varphi_\nu} \qquad (\nu = 2, \ldots k), \tag{3.1.8}$$

then the (R, S) difference procedure is relatively stable with respect to the initial-value problem (3.1.1).

Theorem 3.1.b. If the stabilization parameter L satisfies the inequality

$$L > 2 \cdot \mathrm{Max}\,\frac{-\lambda_\nu \cdot \mathrm{Re}\,A}{1 - \cos\varphi_\nu} \qquad (\nu = 2, \ldots, k), \tag{3.1.9}$$

then the (R, S) difference procedure is absolutely stable with respect to the initial-value problem (3.1.1).

Let us look, in particular, at the case $-\mathrm{Re}\,A < 0$, we see that the growth parameters λ_ν such that $\lambda_\nu - 1 < 0$ are decisive for relative stability whereas all negative λ_ν must be considered for absolute stability. In this case, relative stability implies absolute stability though the converse is not true. In case $-\mathrm{Re}\,A > 0$, / 44 absolute stability implies relative stability.

3.2. Stabilization in the case of a linear differential equation.

The results of the preceding section will now be generalized to the case of a linear (homogeneous) first-order differential equation with variable coefficients. Suppose that we are given the initial-value problem

$$x' = g(t).x, \qquad x(0) = 1. \tag{3.2.1}$$

We assume that the function g(t) is defined and continuous on the interval [0, b] and that g(t) is real-valued in the interval.

The difference equation corresponding to this differential equation is linear and homogeneous. However, the coefficients in it are functions of n:

$$\sum_{\nu=0}^{k} (\tilde{\alpha}_\nu - h\,\tilde{\beta}_\nu\,g_{n+\nu}) \cdot x_{n+\nu} = 0 \qquad (n = 0, 1, \ldots;\ t_n \in [0, b]). \tag{3.2.2}$$

Let us first construct a fundamental system of solutions of the point n = 0. Suppose that v_μ (for $\mu = 1, \ldots, k$) are solutions of the equation

$$R(w, hL) - hg_0 \cdot S(w, hL) = 0 \qquad (g_0 = g(0)).$$

If $g_0 \neq 0$, then, for sufficiently small values of hL, we have $v_\mu \neq v_\nu$ for $\mu \neq \nu$. If $g_0 = 0$, then $v_\mu = w_\mu$, where $R(w_\mu, hL) = 0$ (for $\mu = 1, \ldots, k$) and $w_\mu \neq w_\nu$ (for $\mu \neq \nu$). The determinant

$$\mathrm{Det} \begin{pmatrix} 1 & v_1 & v_1^2 & \cdots & v_1^{k-1} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ 1 & v_k & v_k^2 & \cdots & v_k^{k-1} \end{pmatrix}$$

is, in every case, nonzero for sufficiently small values of hL. Consequently, the sequences $\{v_\mu^n\}$ for ($\mu = 1, \ldots, k$) constitute a fundamental system of solutions for (3.2.2) at the point n = 0. In accordance with Theorems 1.1 and 1.2, the general solution of this difference equation can be written for n = 0 in the form /45

$$x_n = \sum_{\mu=1}^{k} D_\mu \cdot v_\mu^n + \sigma(h) \qquad (3.2.3)$$

under the assumption that hL is chosen in such a way that the quantity $\tilde{\alpha}_0 - hg_0\tilde{\beta}_0$ is nonzero, and this is always possible for sufficiently small hL.

Let us now define

$$z_n^{(\mu)} = w_\mu^{-n} \cdot v_\mu^n \qquad (\mu = 1, \ldots, k) \qquad (3.2.4)$$

with $w\mu \neq 0$ for sufficiently small hL. We then have

$$\sum_{\nu=0}^{k} (\tilde{\alpha}_\nu^{(\mu)} - h\tilde{\beta}_\nu^{(\mu)} \cdot g_{n+\nu}) \cdot z_{n+\nu}^{(\mu)} = 0 \qquad (3.2.5)$$

with

$$\tilde{\alpha}_\nu^{(\mu)} = \tilde{\alpha}_\nu \cdot w_\mu^\nu \qquad (\mu = 1, \ldots, k)$$

$$\tilde{\beta}_\nu^{(\mu)} = \tilde{\beta}_\nu \cdot w_\mu^\nu \qquad (\nu = 0, \ldots, k).$$

For every $\mu = 1, \ldots, k$, we define a new difference procedure possessing the characteristic polynomials

$$R^{(\mu)}(w, hL) = R(w \cdot w_\mu, hL) = \sum_{\nu=0}^{k} \tilde{\alpha}_\nu^{(\mu)} \cdot w^\nu$$

and

$$S^{(\mu)}(w,hL) = S(w \cdot w_\mu, hL) = \sum_{\nu=0}^{k} \tilde{\beta}_\nu^{(\mu)} \cdot w^\nu$$

We can easily verify the following properties:

$$R^{(\mu)}(1,hL) = 0 \qquad (\mu = 1, \ldots, k)$$

and, by application of (3.1.5), /46

$$R^{(\mu)}(1,hL) = w_\mu \cdot R'(w_\mu,hL) = \tilde{\lambda}_\mu^{-1} \cdot S(w_\mu,hL)$$
$$= \tilde{\lambda}_\mu^{-1} \cdot S^{(\mu)}(1,hL).$$

These difference procedures satisfy a consistency condition of the form (2.1.17) if we choose as their second characteristic polynomial the polynomial

$$S_1^{(\mu)}(w,hL) = \tilde{\lambda}_\mu^{-1} \cdot \overline{S^{(\mu)}(w,hL)}$$

Then, equation (3.2.5) becomes

$$\sum_{\nu=0}^{k} (\tilde{\alpha}_\nu^{(\mu)} - h \cdot \tilde{\lambda}_\mu^{-1} \cdot \tilde{\beta}_\nu^{(\mu)} \cdot \bar{\varepsilon}_{n+\nu} \cdot \tilde{\lambda}_\mu) \cdot z_{n+\nu}^{(\mu)} = 0 \qquad (3.2.6)$$
$$(\mu = 1, \ldots, k).$$

In accordance with section 2.3,* the $(R^{(\mu)}, S_1^{(\mu)})$ procedures converge. If we apply these procedures to the initial-value problem

$$z^{(\mu)\prime} = \tilde{\lambda}_\mu \cdot \varepsilon(t) \cdot z^{(\mu)}, \quad z^{(\mu)}(0) = 1, \qquad (3.2.7)$$

whose exact solution is given by

$$z^{(\mu)}(t) = \exp(\tilde{\lambda}_\mu \cdot \int_0^t \varepsilon(\sigma)d\sigma)$$

we get the difference equation (3.2.6). Since the relationship

$$z_\nu^{(\mu)} = w_\mu^{-\nu} \cdot v_\mu^\nu = 1 + O(h)$$

---

*One can show without difficulty that an (R, S) procedure also converges when the zeros of R satisfy not the condition of Theorem 2.3 but the weaker condition $w_\nu = z_\nu (1 + c_\nu \cdot hL + O(h^2))$. The proof is analogous to that given in section 2.3.

holds for the initial values $Z_\nu^{(\mu)}$ (for $\nu = 0, 1, \ldots, k-1$ and $\mu = 1, \ldots, k$), we have /47

$$z_n^{(\mu)} = z^{(\mu)}(t_n) + O(h) .$$

Therefore, the asymptotic behavior of the solution of (3.2.2) as $h \to 0$ (with $nh = t_n$ and L fixed) is given by

$$x_n = \sum_{\mu=1}^{k} D_\mu \cdot \varphi_\mu^n \cdot z^{(\mu)}(t_n) + O(h)$$

Since $\tilde{\lambda}_\mu = \tilde{\lambda}_\mu + O(h)$, we then get

$$x_n = D_1 \exp\left(\int_0^{t_n} g(t)\,dt\right) + \sum_{\mu=2}^{k} D_\mu \cdot \exp(\ln \varphi_\mu) \cdot \exp\left(\frac{L}{2} \cdot \frac{1-z_\mu}{z_\mu} t_n\right) + \lambda_\mu \cdot \int_0^{t_n} g(t)dt + O(h).$$

Again, we see that the first summand approximates the exact solution of our initial-value problem.

For the solution x_n to be relatively stable, the inequality

$$\frac{L}{2} \operatorname{Re} \frac{1-z_\mu}{z_\mu} \cdot t_n + \lambda_\mu \cdot \int_0^{t_n} g(t)dt < \int_0^{t_n} g(t)dt$$

must, in accordance with the definition given in section 3.1 (and appropriately modified) hold for all $\mu = 2, \ldots, k$.

Theorem 3.2.a. The (R, S) difference procedure is relatively stable with respect to the initial-value problem (3.2.1) if the stabilization parameter L satisfies the inequality

$$L > \frac{2}{t_n} \cdot \max_{\nu=2,\ldots,k} \frac{(\lambda_\nu - 1) \cdot \int_0^{t_n} g(t)dt}{1 - \cos \psi_\nu}, \quad t_n \in [0, b], \tag{3.2.10}$$

(Translator's note: No equations are numbered (3.2.8) or (3.2.9) in the original German).

In a similar manner, we obtain the condition for absolute stability: /48

Theorem 3.2.b. The (R, S) difference procedure is absolutely stable with respect to (3.2.1) if the stabilization parameter L satisfies the inequality

$$L > \frac{2}{t_n} \cdot \max_{\nu=2,\ldots,k} \frac{\lambda_\nu \cdot \int_0^{t_n} g(t)dt}{1-\cos\varphi_\nu} , \quad t_n \in [0,b] , \tag{3.2.11}$$

One should note that all these theorems describe a definite asymptotic behavior as $h \to 0$ with $nh = t_n$ fixed and L fixed.

In particular, if we set $g(t) = -\mathrm{Re}A$, where A is a nonzero complex number, inequalities (3.2.10) and (3.2.11) becomes inequalities (3.1.8) and (3.1.9).

The distinction with section 3.1 obviously consists in the fact that in the present section the stabilization parameter L has to satisfy an inequality whose right-hand member does not remain constant. In general, with every step, one has to stabilize anew; that is, one works with a variable L. However, since hL must belong to the interval (0, 2), one first finds the maximum value of the right-hand member of (3.2.10) (or of (3.2.11)) (for a given interval of integration [0, b]). The step h must then be chosen in such a way that the inequality $0 \leq hL \leq 2$ remains satisfied for all $t_n \in [0, b]$.

3.3. Stabilization in the case of a nonlinear differential equation.

In this section, we shall investigate the numerical integration of the initial-value problem (1.1.1, 2) with the help of the (R, S) procedure. Of course, we assume here that the function f(t, x) is at least (k + 3) times continuously differentiable with respect to x in the interval [0, b].

In contrast with the preceding, we shall now discuss directly the asymptotic behavior of the discretization error

$$e_n = x_n - x(t_n) .$$

Let us assume that the values x_n exactly satisfy the difference equation

$$\sum_{\nu=0}^{k} \tilde{\alpha}_\nu \, x_{n+\nu} - h \sum_{\nu=0}^{k} \tilde{\beta}_\nu f(t_{n+\nu} , x_{n+\nu}) = 0 \tag{3.3.1}$$

Let us subtract from (3.3.1) the corresponding expression for the difference operator M:

$$\sum_{\nu=0}^{k} \left(\tilde{\alpha}_\nu \, x(t_{n+\nu}) - h \tilde{\beta}_\nu \, f(t_{n+\nu} , x(t_{n+\nu})) \right) = -F(t_n) \cdot h^{k+3} + O(h^{k+4}) ,$$

where

$$-F(t_n) = C_{k+3} \cdot x^{(k+3)}(t_n) + \frac{L}{2} \cdot C^{*}_{k+2} \cdot x^{(k+2)}(t_n) ,$$

We get

$$\sum_{\nu=0}^{k} \tilde{\alpha}_\nu \, e_{n+\nu} - h \sum_{\nu=0}^{k} \tilde{\beta}_\nu \cdot \left[f(t_{n+\nu}, x_{n+\nu}) - f(t_{n+\nu}, x(t_{n+\nu})) \right] =$$

$$= F(t_n) \cdot h^{k+3} + O(h^{k+4}),$$

or, since $x_n = e_n + x(t_n)$,

$$\sum_{\nu=0}^{k} \tilde{\alpha}_\nu \, e_{n+\nu} - h \sum_{\nu=0}^{k} \tilde{\beta}_\nu \left(f_x(t_{n+\nu}, x_{n+\nu}) \cdot e_{n+\nu} + O(e_{n+\nu}^2) \right) =$$

$$= F(t_n) \cdot h^{k+3} + O(h^{k+4}) . \qquad (3.3.2)$$

(Here, we have assumed that the function f(t, x) has a continuous first partial derivative with respect to x.)

Suppose that the initial values for e_n are sufficiently often differentiable functions of h:

$$e_\mu = \delta_\mu(h) \quad \text{with} \quad \delta_\mu(h) = O(h^q), \quad q \geq 1$$
$$(\mu = 0, 1, \ldots, k-1) .$$

Then, the relation / 50

$$e_n = O(h^r), \qquad r = \min(k+2, q) .$$

(see [20], Theorem 5.11, p. 248) holds for the discretization of e_n. For what follows, it will be convenient to use not e_n but the "augmented" discretization error

$$E_n = h^{-r} \cdot e_n$$

Then, E_n satisfies the difference equation

$$\sum_{\nu=0}^{k} \tilde{\alpha}_\nu \, E_{n+\nu} - h \sum_{\nu=0}^{k} \tilde{\beta}_\nu \cdot f_{x,n+\nu} \cdot E_{n+\nu} = F(t_n) \cdot h^{k+3-r} + O(h^2) , \qquad (3.3.3)$$

since $r + 1 \geq 2$ and $k + 4 - r \geq 2$. Also, $f_{x, n+\nu}$ stands for $f_x(t_{n+\nu}, x_{n+\nu})$. The corresponding initial values are

$$E_\mu = \Delta_\mu(h) = h^{-r} \cdot \delta_\mu(h) \qquad (\mu = 0, 1, \ldots, k-1) . \qquad (3.3.4)$$

We know from the theory of difference equations that the solution of (3.3.3) is the sum of the solution of the homogeneous difference equation

$$\sum_{\nu=0}^{k} (\tilde{\alpha}_\nu - h \tilde{\beta}_\nu \cdot f_{x,n+\nu}) \cdot E_{n+\nu} = 0 \qquad (3.3.5)$$

with initial values (3.3.4) and a solution of the nonhomogeneous equation with initial values 0, ..., 0. We are interested in the nature of the solution of the homogeneous equation with the initial-values mentioned since this equation is the key to the discussion of the stability. We spoke in detail of a difference equation of the form (3.3.5), so that here we will confine ourselves to proving the following result: the solution of (3.3.5) determined by the initial values (3.3.4) satisfies the asymptotic relationship

$$E_n = \sum_{\mu=1}^{k} A_\mu \cdot \varrho_\mu^n \cdot Z^{(\mu)}(t_n) + G(h) \quad (h \to 0,\ nh = t_n,\ L \text{ fixed}),$$

where $Z^{(\mu)}(t)$ is the exact solution of the initial-value problem / 51

$$Z^{(\mu)'} = \tilde{\lambda}_\mu \cdot f_x(t) \cdot Z^{(\mu)},\ Z^{(\mu)}(0) = 1 \qquad (\mu = 1, \ldots, k) \qquad (3.3.6)$$

It has the form

$$Z^{(\mu)}(t) = \exp(\tilde{\lambda}_\mu \int_0^t f_x(\sigma) : d\sigma) \qquad f_x(t) = f_x(t, x(t)) .$$

The relationship for the error E_n is written

$$E_n = \sum_{\mu=1}^{k} A_\mu \cdot \exp(\ln \varrho_\mu) \cdot \exp(\frac{L}{h} \cdot \frac{1-z_\mu}{z_\mu} t_n + \lambda_\mu \cdot \int_0^{t_n} f_x(t)dt) + G(h) \qquad (3.3.7)$$

as $h \to 0$ with $nh = t_n$ and L fixed. The inequalities for the stabilization parameter, which follow immediately from this, are contained in the following two theorems:

Theorem 3.3.a. The (R, S) difference procedure is relatively stable with respect to the initial-value problem (1.1.1, 2), where f(t, x) satisfies the assumptions made at the beginning of this section provided the stabilization parameter L satisfies the inequality

$$L > \frac{2}{t_n} \max_{\mu=2,\ldots,k} \frac{(\lambda_\mu - \lambda_1) \cdot \int_0^{t_n} f_x(t)dt}{1-\cos\varphi_\mu} \quad (t_n \in [0,b]) \qquad (3.3.8)$$

Theorem 3.3.b. The (R, S) difference procedure is absolutely stable with respect to the initial-value problem (1.1.1, 2) under the additional assumptions of the present section provided the stabilization parameter L satisfies the inequality

$$L > \frac{2}{t_n} \cdot \max_{\mu=2,\ldots,k} \frac{\lambda_n \cdot \int_0^{t_n} f_x(t)dt}{1-\cos\varphi_\mu} \quad (t_n \in [0,b]) \qquad (3.3.9)$$

If we compare these two results with the inequalities of Theorems 3.2 and 3.1, we notice immediately that the earlier relationships are special cases of the present general stabilization inequalities since in those theorems the function $f(t, x) = g(t) \cdot x$ and the function $f(t, x) = -Az$ satisfy respectively the additional assumptions. /52

CHAPTER 4

SYSTEMS OF DIFFERENTIAL EQUATIONS

4.1. Systems of linear equations with constant coefficients.

The purpose of the present chapter is to extend the results proven in the last chapter to systems of first-order linear and nonlinear differential equations.

First of all, let us again investigate the simple case of a system with constant coefficients, namely,

$$X' = AX, \quad X(0) = S \qquad \left(' = \frac{d}{dt} \right), \tag{4.1.1}$$

where

$$X = X(t) = (x^1(t), \ldots, x^m(t))^T,$$

$$A = (a_{ij}) \quad \text{(square, of order m).}$$

Suppose that the elements of A are nonzero real numbers. Let us denote the eigenvalues of A by $s_1, \ldots, s_m$.

If the eigenvalues of A are all simple, then there exists a regular matrix T such that the matrix $D = T^{-1}AT$ is a diagonal matrix. The elements on the principal diagonal of D are the eigenvalues of A since these are invariant under a similarity transformation. Let us define the vector Y by $Y = T^{-1}X$. It follows that

$$Y' = T^{-1}X' = T^{-1}ATY = DY.$$

The transformed system is then

$$(Y^\nu)' = s_\nu \cdot Y^\nu, \quad Y(0) = T^{-1} \cdot X(0) \qquad (\nu = 1, \ldots, m).$$

The m differential equations are of the form (3.1.1). There are two ways open to us: we can stabilize each equation individually or we can choose L in such a way that we achieve relative or absolute stability simultaneously for all m equations.

The first possibility leads, for relative stability, in accordance with (3.1.8) to the equality

/54

$$L_\mu > 2 \cdot \operatorname{Max}_{\nu=2,\ldots,k} \frac{(\lambda_\nu - 1) \cdot \operatorname{Re} a_\mu}{1-\cos \varphi_\nu}, \quad (\mu = 1, \ldots, m), \tag{4.1.2}$$

whereas, for absolute stability, we get the inequality

$$L_\mu > 2 \cdot \operatorname{Max}_{\nu=2,\ldots,k} \frac{\lambda_\nu \cdot \operatorname{Re} a_\mu}{1-\cos \varphi_\nu}, \quad (\mu = 1, \ldots, m) \tag{4.1.3}$$

in accordance with (3.1.9).

If we wish to use a single value of L to stabilize all m equations together (which is primarily advantagenous in the case of a small spectral radius of A), L must satisfy the inequalities

$$L > 2 \cdot \operatorname{Max}_{\substack{\nu=2,\ldots,k \\ \mu=1,\ldots,m}} \frac{(\lambda_\nu - 1) \cdot \operatorname{Re} a_\mu}{1-\cos \varphi_\nu}$$

for relative stability and it must satisfy the inequalities

$$L > 2 \cdot \operatorname{Max}_{\substack{\nu=2,\ldots,k \\ \mu=1,\ldots,m}} \frac{\lambda_\nu \cdot \operatorname{Re} a_\mu}{1-\cos \varphi_\nu}$$

for absolute stability. However, the matrix A will not in the general case have only simple eigenvalues. In this case, we use the following theorem of Schur ([2], p. 195) (treated in linear algebra) in our further discussion:

Corresponding to a given matrix A there always exists a unit matrix U such that the matrix $T = U^* A U$ is a (lower triangular matrix. The elements on the principal diagonal of T are the eigenvalues of A.

Therefore, the transformed system (4.1.1) is written

$$Y' = T.Y, \quad Y(0) = U^* \beta, \tag{4.1.4}$$

since U is a matrix with constant coefficients. Specifically, we have

/55

$$(Y^\mu)' = a_\mu \cdot Y^\mu + \sum_{\nu=1}^{\mu-1} b_{\mu\nu} Y^\nu, \qquad \mu = 1, \ldots, m,$$

where the first equation has the simple form

$$(Y^1)' = a_1 \cdot Y^1$$

The difference equation corresponding to the μth equation of this system is non-homogeneous. It is

$$\sum_{\nu=0}^{k} \tilde{\alpha}_\nu \; Y^\mu_{n+\nu} \;=\; a_\mu \cdot h \cdot \sum_{\nu=0}^{k} \tilde{\beta}_\nu \; Y^\mu_{n+\nu} \;+\; h \sum_{\nu=0}^{k} \tilde{\beta}_\nu \; \psi^\mu_{n+\nu} \qquad (4.1.5)$$

with the known function

$$\psi^\mu_{n+\nu} \;=\; \sum_{j=1}^{\mu-1} b_{\mu j} \cdot Y^j_{n+\nu} \;, \qquad \psi^1_{n+\nu} \;=\; 0 \;.$$

As was mentioned earlier, the general solution of (4.1.5) is the sum of the general solution of the homogeneous equation

$$\sum_{\nu=0}^{k} (\tilde{\alpha}_\nu \;-\; a_\mu \cdot h \; \tilde{\beta}_\nu) \; Y^\mu_{n+\nu} \;=\; 0 \qquad (4.1.6)$$

and a particular solution of the nonhomogeneous equation (4.1.5). Consequently, the eigenvalues of the matrix A are again decisive for relative and absolute stability although they are no longer necessarily destinct from each other. This fact has no effect on the inequalities (4.1.2) and (4.1.3), so that these are in general valid for the initial-value problem (4.1.1).

4.2. Systems of linear differential equations with variable coefficients. /56

Suppose that we are given the system

$$x' \;=\; A(t).X \;, \qquad X(0) \;=\; E \;=\; (1, \; \ldots \; , \; 1)^T \;. \qquad (4.2.1)$$

Now, the elements $a_{ij} = a_{ij}(t)$ of the square matrix A(t) (of order m) are functions of t, and we assume that these functions are defined, real-valued, and continuous on the finite interval [0, b]. Furthermore, we assume that $A(t) \neq 0$ for $t \in [0, b]$.

We again apply the (R, S) difference procedure for the numerical integration of this initial-value problem. The corresponding difference equation is then

$$\sum_{\nu=0}^{k} \tilde{\alpha}_\nu \; X_{n+\nu} \;=\; h \cdot \sum_{\nu=0}^{k} \tilde{\beta}_\nu \; A_{n+\nu} \; X_{n+\nu} \;. \qquad (4.2.2)$$

Here, $A_\mu = A(t_\mu) = (a_{ij}(t_\mu))$. Just as in the scalar case, it is a question of a linear homogeneous difference equation with variable coefficients.

For our subsequent discussion, we use the following notation:

$$X_\mu \;=\; \begin{pmatrix} x^1_\mu \\ \vdots \\ x^m_\mu \end{pmatrix} \qquad X(t_\mu) \;=\; \begin{pmatrix} x^1(t_\mu) \\ \vdots \\ x^m(t_\mu) \end{pmatrix}$$

Just as before, $X(t_r)$ stands for the exact solution of (4.2.1) at the point $t = t_r$, whereas X_r denotes an approximation for this value at the point $t = t_r$. I denotes the unit matrix of order m.

To generalize the results found for the scalar case, the procedure chosen in section 3.2 cannot be used in its entirety. The reason is that, in contrast with the initial-value problem (3.2.1), the exact solution of (4.2.1) cannot be written out explicity. This means that we cannot solve the initial-value problem corresponding to (3.2.7) and that we must therefore seek a way of at least estimating this solution.

To begin with, we proceed much as in section 3.2. Instead of (4.2.2), we write / 57

$$\sum_{\nu=0}^{k} (\tilde{\alpha}_\nu I - h\tilde{\beta}_\nu A_{n+\nu}) X_{n+\nu} = 0 . \qquad (4.2.3)$$

With appropriate modification, the relations (3.2.3)-(3.2.6) carry over immediately to the vector case. Finally, as in section 3.2, we have, for the solution of the difference equation (4.2.3), the asymptotic relationship (as $h \to 0$ with $nh = t_n$ and L fixed)

$$X_n = \sum_{\mu=1}^{k} Z^{(\mu)}(t_n) \cdot w_\mu^n \cdot D_\mu + O(h) , \qquad (4.2.4)$$

where the matrix $Z^{(\mu)}(t)$ is the exact solution of the system

$$Z^{(\mu)\prime} = \tilde{\lambda}_\mu \cdot A(t) \cdot Z^{(\mu)} , \quad Z^{(\mu)}(0) = I \qquad (\mu = 1, \ldots, k) \qquad (4.2.5)$$

As was mentioned at the beginning of the present section, the solution of this system cannot be expressed explicity. On the other hand, it is possible to find an estimate for it.

To this end, we first introduce a vector norm and a corresponding subordinate matrix norm. Let $\| X \|$ denote the norm of the vector X (in some suitable sense). Then, we choose as the matrix norm

$$A = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} .$$

Such a norm possessesses the following familiar properties:

a) $\|A\| \geq 0$, $\|A\| = 0 \iff A = 0$.
b) $\|cA\| = |c| \cdot \|A\|$
c) $\|A + B\| \leq \|A\| + \|B\|$
d) $\|AB\| \leq \|A\| \cdot \|B\|$.

Furthermore, $\|A\,x\| \leq \|A\| \cdot \|x\|$.

It is known (see, for example, [6]) that the solution of the system (4.2.1) can be estimated by

$$\|x(t)\| \leq \|x(0)\| \cdot \exp\left(\int_0^t \|A(\sigma)\|\, d\sigma\right) \qquad /58$$

In [11] (p. 10), Dahlquist gave a sharper estimate. This rests on the concept of a "logarithmic norm" or a matrix which he introduced. This special norm (actually a pseudo-norm as we shall soon see) is defined by

$$\mu[A] = \lim_{h \to 0+} \frac{\|I + hA\| - 1}{h} .$$

It has, among other properties, the following:

1) $\mu[A]$ can, in contrast with $\|A\|$ be negative
2) $|\mu[A]| \leq \|A\|$.
3) $\mu[cA] = c \cdot \mu[A]$, $\quad c > 0.$
4) $\mu[A + B] \leq \mu[A] + \mu[B]$. (4.2.6)

With the help of this form, Dahlquist then found the following estimate for the solution of (4.2.1):

$$\|x(t)\| \leq \|x(0)\| \cdot \exp\left(\int_0^t \mu[A(\sigma)]\, d\sigma\right) ,$$

which, by virtue of (4.2.6), represents in general a sharpening of the above-cited estimate.

For our subsequent discussion, let us assume that the basic vector norm is the Euclidean norm

$$\|y\| = \left(\sum_{\nu=1}^{m} |y^{\nu}|^2\right)^{1/2}$$

Then, for the subordinate matrix norm

$$\|A\| = \Lambda_{\max}^{1/2}$$

where $\Lambda_{\max}$ is the greatest eigenvalue of the matrix $A^* A$ (where A^* denotes the adjoint or complex-conjugate transpose of A).

In this case, μ [A] is equal to the greatest eigenvalue of the (symmetric or Hermitian) matrix (A* + A) / 2. / 59

Let μ' [A] denote the smallest eigenvalue of the matrix (A* + A) / 2. Then,

$$\mu[cA] = c \cdot \mu'[A], \qquad c > 0$$

and

$$\mu[cA] = c \cdot \mu'[A], \qquad c < 0,$$

since, for $c < 0$,

$$\mu[cA] = |c| \cdot \mu[-A] = -c \cdot \mu'[A] = c \cdot \mu'[A]$$

Since the matrix B (t) = (A* (t) - A (t)) / 2 is symmetric (we are assuming that A (t) is real-valued), the following relation holds for the eigenvalues Λ_ν (ε) of B (t) + $\varepsilon \cdot$ B (t):

$$\Lambda_\nu(\varepsilon) = \Lambda_\nu(0) + O(\varepsilon) \qquad (\varepsilon \to 0)$$

(see, for example [46], [47]). Thus, we find, under observation of (3.1.6),

$$\mu[\tilde{\lambda}_\mu \cdot A(t)] = \mu[\lambda_\mu \cdot A(t)] + O(h) \qquad (h \to 0,\ \mu = 1, \ldots, k)$$

Furthermore, just as before,

$$|w_\mu|^n = \exp\left(\frac{L}{2} \cdot (\cos\varphi_\mu - 1) \cdot t_n\right) + O(h),$$

as $h \to 0$ with $nh = t_n$ and L fixed. From these results, we finally obtain the sought-for estimate for the solution of the initial-value problem (4.2.5):

$$\| z^{(\mu)}(t) \| \le \| z^{(\mu)}(0) \| \cdot \exp\left(\lambda_\mu \cdot \int_0^t \mu[A(\sigma)]\, d\sigma\right) + O(h)$$

for all μ such that $\lambda_\mu > 0$ and

$$\| z^{(\mu)}(t) \| \le \| z^{(\mu)}(0) \| \cdot \exp\left(\lambda_\mu \cdot \int_0^t \mu'[A(\sigma)]\, d\sigma\right) + O(h)$$

for those μ such that $\lambda_\mu < 0$. Substitution into (4.2.4) yields / 60

$$\| x_n \| \le \sum_{\{\mu \mid \lambda_\mu > 0\}} c_\mu \cdot \exp\left(\frac{L}{2}(\cos\varphi_\mu - 1)t_n + \lambda_\mu \int_0^{t_n} \mu[A(t)]\, dt\right)$$

$$+\sum_{\{\mu\,|\,\lambda_\mu<0\}} G_\mu \cdot \exp\left(\frac{L}{2}\cdot(\cos\varphi_\mu - 1)\,t_n + \lambda_\mu \int_0^{t_n} \mu'[A(t)]\,dt\right) + O(h)$$

$$(t_n \in [0, b])$$

where

$$G_\mu = |D_\mu| \cdot \|z^{(n)}(0)\| \quad , \qquad \mu = 1, \ldots, k \, .$$

The stabilization inequalities for L can now be written immediately if, just as in section 3.2, we carry over in a logical way the definitions of relative and absolute stability. If we set

$$U_1 = \frac{2}{t_n} \max_{\{\mu\,|\,\lambda_\mu>0,\ \mu\neq 1\}} \frac{(\lambda_\mu - 1)\cdot\int_0^{t_n} \mu\,[A(t)]\,dt}{1-\cos\varphi_\mu}$$

$$U_2 = \frac{2}{t_n} \max_{\{\mu\,|\,\lambda_\mu<0\}} \frac{(\lambda_\mu - 1)\int_0^{t_n} \mu'[A(t)]\,dt}{1-\cos\varphi_\mu}$$

$$V_1 = \frac{2}{t_n} \cdot \max_{\{\mu\,|\,\lambda_\mu>0,\ \mu\neq 1\}} \frac{\lambda_\mu \cdot \int_0^{t_n} \mu\,[A(t)]\,dt}{1-\cos\varphi_\mu}$$

$$V_2 = \frac{2}{t_n} \cdot \max_{\{\mu\,|\,\lambda_\mu<0\}} \frac{\lambda_\mu \cdot \int_0^{t_n} \mu'[A(t)]\,dt}{1-\cos\varphi_\mu}$$

we obtain the following:

Theorem 4.1.a. The (R, S) difference procedure is relatively stable with respect to the initial-value problem (4.2.1) if the stabilization parameter L satisfies the inequality

$$L > \operatorname{Max}(U_1, U_2) \tag{4.2.7}$$

Theorem 4.1b. The (R, S) difference procedure is absolutely stable with respect to the initial-value problem (4.2.1) if L satisfies the inequality

/61

$$L > \operatorname{Max}(V_1, V_2) \tag{4.2.8}$$

For the scalar case $A(t) = g(t)$, we have $\mu[g(t)] = \mu'[g(t)]$, and the above-given inequalities become the inequalities found earlier for the scalar case (3.2.10) and (3.2.11).

4.3. Systems of nonlinear differential equations.

In conclusion, we shall discuss the case of a system of m nonlinear first-order differential equations:

$$X' = F(t,X) ; \quad X(0) = \alpha, \tag{4.3.1}$$

where

$$F(t,X) = \begin{pmatrix} F^1(t,X) \\ \vdots \\ F^m(t,X) \end{pmatrix} \qquad X = X(t) = \begin{pmatrix} X^1(t) \\ \vdots \\ X^m(t) \end{pmatrix}$$

Suppose that the elements of F (t, X) are, for $t \in [0, b]$, continuous and real-valued functions. If the elements of the Jacobian matrix

$$F_X(t,X) = \left(\frac{\partial F^i(t,X)}{\partial X^j}\right) \quad (i,j = 1, \ldots, m)$$

exist for $t \in [0, b]$ and are continuous and continuously differentiable functions of t in that interval, then (4.3.1) has a unique continuously differentiable solution X (t) in that interval.

The relationships that are needed for the derivation of the stabilization equations are all contained in sections 3.3 and 4.2. Therefore, we shall not repeat them. We find that the "augmented" discretization error

$$E_n = h^{-r} \cdot (X_n - X(t_n))$$

satisfies the asymptotic relationship (see (3.3.3)).

/62

$$\|E_n\| \leq \sum_{\{\mu \mid \lambda_\mu > 0\}} G'_\mu \cdot \exp\left(\frac{L}{2} \cdot (\cos\varphi_\mu - 1) \cdot t_n + \lambda_\mu \cdot \int_0^{t_n} \mu\,[F_X(t)]\,dt\right)$$

$$+ \sum_{\{\mu \mid \lambda_\mu < 0\}} G'_\mu \cdot \exp\left(\frac{L}{2} (\cos\varphi_\mu - 1) \cdot t_n + \lambda_\mu \cdot \int_0^{t_n} \mu'\,[F_X(t)]\,dt\right) + O(h)$$

$$(t_n \in [0,b])$$

where $F_X(t) = F_X(t, X(t))$ and the quantities G'_μ are arbitrary constants determined by the initial values of E_n. Let us define

$$U_3 = \frac{2}{t_n} \cdot \max_{\{\mu \mid \lambda_\mu > 0,\ \mu \neq 1\}} \frac{(\lambda_\mu - 1) \cdot \int_0^{t_n} \mu \left[F_x(t)\right] dt}{1 - \cos \varphi_\mu}$$

$$U_4 = \frac{2}{t_n} \cdot \max_{\{\mu \mid \lambda_\mu < 0\}} \frac{(\lambda_\mu - 1) \cdot \int_0^{t_n} \mu' \left[F_x(t)\right] dt}{1 - \cos \varphi_\mu}$$

$$V_3 = \frac{2}{t_n} \cdot \max_{\{\mu \mid \lambda_\mu > 0,\ \mu \neq 1\}} \frac{\lambda_\mu \cdot \int_0^{t_n} \mu \left[F_x(t)\right] dt}{1 - \cos \varphi_\mu}$$

$$V_4 = \frac{2}{t_n} \cdot \max_{\{\mu \mid \lambda_\mu < 0\}} \frac{\lambda_\mu \cdot \int_0^{t_n} \mu' \left[F_x(t)\right] dt}{1 - \cos \varphi_\mu}$$

$$(t_n \in [0, b])$$

With this notation, we have the following results:

Theorem 4.2.a. The (R, S) difference procedure is relatively stable with respect to the initial-value problem (4.3.1) if the stabilization parameter L satisfies the inequality

$$L > \max (U_3, U_4) \qquad (4.3.2)$$

Theorem 4.2.b. The (R, S) difference procedure is absolutely stable with respect to the initial-value problem (4.3.1) if the stabilization parameter L satisfies the inequality / 63

$$L > \max (V_3, V_4) \qquad (4.3.3)$$

We see immediately that the inequalities, for $F(t, X) = A(t) \cdot X$ lead immediately to the relationships (4.2.7) and (4.2.8).

CHAPTER 5

/64

CONDITIONAL STABILITY IN THE CASE OF AN EXPLICIT PROCEDURE OF MAXIMUM ORDER

Every difference procedure whose first characteristic polynomial has, in addition to the zero $z = 1$, other zeros on the unit circle can exhibit conditional stability, so that this is not true merely of the optimal procedure described. In particular, this is true for those difference procedures that have the same first characteristic polynomials as the optimum procedures whereas the other characteristic polynomial is of degree k-1. Such a procedure is called explicit. A familiar example of this is the procedure

$$x_{n+2} - x_n = 2h.f_{n+1} \qquad \text{(Nyström, } q = 0\text{)}, \tag{5.1}$$

which rests on the midpoint rule and whose left side coincides with the Milne-Simpson procedure. The zeros of $\rho(z)$ are $z_1 = 1$ and $z_2 = -1$, and the corresponding growth parameters are $\lambda_1 = 1$ and $\lambda_2 = -1$.

By using the results of Chapter II, it is easy to find a modified procedure corresponding to an explicit (ρ, σ) difference procedure with

$$\rho(z) = \sum_{\nu=0}^{k} \alpha_\nu z^\nu$$
$$\sigma(z) = \sum_{\nu=0}^{k-1} \beta_\nu z^\nu,$$

where $\rho(z)$ satisfies the assumption of an optimal procedure. Suppose that the first characteristic polynomial of this procedure is again R (w, hL) defined in (2.1.6). In analogy with (1.3.8), we then set up the series

$$\frac{\rho(w)}{\log w} = \sum_{\nu=0}^{\infty} c_\nu (w-1)^\nu,$$

and

/65

$$\frac{\rho'(w).(w-1)}{\log w} = \sum_{\nu=0}^{\infty} \gamma_\nu (w-1)^\nu \qquad (\log 1 = 0),$$

We define

$$\sigma(w) = \sum_{\nu=0}^{k-1} c_\nu \cdot (w-1)^\nu \quad , \quad \sigma^\circ(w) = \sum_{\nu=0}^{k-1} \gamma_\nu \cdot (w-1)^\nu .$$

The second modified polynomial is then defined by

$$S(w, hL) = \sigma(w) + \frac{hL}{2} \cdot \sigma^\circ(w)$$

Obviously, this (R, S) procedure is an explicit procedure and it satisfies the stability condition of Theorem 2.3 and the consistency conditions of Theorem 2.4.

Theorem 5.1. The order of the explicit (R, S) procedure is equal to k as long as the polynomial $\rho(z)$ satisfies the assumptions of section 2.1. The error constant of this procedure is the same as that of the explicit (ρ, σ) difference procedure.

Proof: The first part of the theorem is an immediate consequence of a theorem proven by Dahlquist ([10], p. 52, Theorem 5), which asserts that the order of a stable and consistent difference procedure cannot exceed its degree, and of Lemma 2.1. In accordance with this lemma, the difference operator corresponding to the polynomials $\rho^*(w)$ and $\sigma^*(w)$ are, according to the construction of $\sigma^*(w)$, of order at least k. By virtue of the definition of order, it follows that

$$L[x(t), h] = C_{k+1} \cdot x^{(k+1)}(t) \cdot h^{k+1} + O(h^{k+2}) \quad (C_{k+1} \neq 0) ,$$

Furthermore,

$$M^\circ[x(t); h] = C^\circ_{k+1} \cdot x^{(k+1)}(t) \cdot h^{k+1} + O(h^{k+2}) \quad (C^\circ_{k+1} \neq 0) ,$$

With

$$M = L + \frac{hL}{2} \cdot M^\circ$$

we have

$$M[x(t); hL] = C_{k+1} x^{(k+1)}(\cdot) \cdot h^{k+1} + O(h^{k+2}) . \qquad (5.2)$$

/ 66

We see that the modified difference operator has also the maximum possible order p = k.

Furthermore, we also see from (5.2) that the error constant of the modified (R, S) procedure is equal to that of the given (ρ, σ) procedure, namely,

$$C = \frac{1}{\rho'(1)} \cdot \left[\frac{1}{(k+1)!} \sum_{\nu=1}^{k} \nu^{k+1} \cdot \alpha_\nu - \frac{1}{k!} \sum_{\nu=1}^{k-1} \nu^k \cdot \beta_\nu \right].$$

We refrain from a discussion of conditional stability and the corresponding stabilization inequalities since the results of Chapters III and IV can be carried over immediately to the explicit case.

In contrast, in Chapter 7, we shall introduce a few special explicit difference procedures, namely, the modified procedure of Nyström ($q = 0$: the middle point rule), three procedures with $k = 4$, and one procedure with $k = 6$.

CHAPTER 6

OPTIMAL PROCEDURES FOR k = 2, k = 4, AND k = 6

6.1. The modified procedure of Milne-Simpson.

For k = 2, there exists only one optimal procedure with real coefficients, namely, the Milne-Simpson procedure. Its characteristic polynomials are

$$\rho(z) = z^2 - 1$$
$$\sigma(z) = \frac{1}{3}(z^2 + 4z + 1).$$

The polynomial $\rho(z)$ has zeros $z_1 = 1$ and $z_1 = -1$; the growth parameters of the procedure are $\lambda_1 = 1$ and $\lambda_1 = -1/3$.

For the modified procedure, we find, in accordance with Chapter 2,

$$R(w,hL) = w^2(1 + hL) - hL.w - 1$$
$$S(w,hL) = \frac{1}{12}\left[(4 + 5hL)w^2 + (16 + 8hL)w + 4 - hL\right].$$

The error constant of the (ρ, σ) procedure has the known value

$$C = -\frac{1}{180},$$

whereas the error constant introduced by the polynomials $\rho^*(w)$ and $\sigma^*(w)$ are found to be

$$C^* = -\frac{1}{90}$$

6.2. Optimal procedures with k = 4.

Let us now look at those optimal procedures whose characteristic polynomials $\rho(z)$ have real integer coefficients. It is known [20] (p. 280) that three such procedures exist:

I.
$$\rho(z) = z^4 - 1$$
$$\sigma(z) = \frac{1}{45}(14z^4 + 64z^3 + 24z^2 + 64z + 14).$$

The zeros of $\rho(z)$ and the growth parameters of this procedure are

/68

$$z_1 = 1, \quad z_2 = \bar{z}_4 = \exp(i\,\frac{\pi}{3}),$$

$$\lambda_1 = 1, \quad \lambda_2 = \lambda_4 = \frac{1}{15}, \quad \lambda_3 = -\frac{10}{15}.$$

For the supplementary polynomials $\rho^*(z)$ and $\sigma^*(z)$, we find

$$\rho^*(z) = 4z^4 - 4z^3$$

$$\sigma^*(z) = \frac{1}{100}\,(251z^4 + 660z^3 - 204\,z^2 + 100\,z - 10).$$

Thus, we have, for the modified characteristic polynomials,

$$R(w,hL) = w^4(1 + 2hL) - 2hL \cdot w^3 - 1$$

$$S(w,hL) = \frac{1}{360}\Big[(112 + 251.hL)w^4 + (512 + 640.hL)w^3 + (100 - 204.hL)w^2 + (512 + 100.hL)w + 112 - 10hL\Big].$$

The two error constants have the values

$$C = -\frac{2}{945} \quad \text{and} \quad C^* = -\frac{3L}{320}.$$

II.

$$\rho(z) = z^4 + z^3 - z - 1$$

$$\sigma(z) = \frac{3}{10} \cdot (z^4 + 6z^3 + 6z^2 + 6z + 1).$$

Here, zeros of $\rho(z)$ and growth parameters are

$$z_1 = 1, \quad z_2 = \bar{z}_4 = \exp(i\,\frac{2\pi}{3}), \quad z_3 = -1,$$

$$\lambda_1 = 1, \quad \lambda_2 = \lambda_4 = \frac{1}{10}, \quad \lambda_3 = -\frac{3}{5}.$$

The modified procedure has the supplementary polynomials

$$\rho^*(z) = 4z^4 - z^3 - 3z^2 - z + 1$$

$$\sigma^*(z) = \frac{1}{120}\,(161z^4 + 503z^3 + 63z^2 - 74z - 40).$$

Consequently, we have, for the modified characteristic polynomials,

/69

$$R(w,hL) = w^4 \cdot (1+2hL) + w^3(1-\frac{hL}{2}) - \frac{3hL}{2} \cdot w^2 - w(1 + \frac{hL}{2}) + \frac{hL}{2} - 1$$

$$S(w,hL) = \frac{1}{240}\Big[(72 + 161hL)w^4 + (432 + 560hL)w^3 + (432 + 90hL)w^2 + (432 - 74hL)w + 72 - 40hL\Big].$$

For the error constants, we get the values

$$C = -\frac{1}{840} \quad \text{and} \quad C^* = -\frac{177}{2800}.$$

III.

$$\rho(z) = z^4 - z^3 + z - 1$$

$$\sigma(z) = \frac{1}{90}(29z^4 + 84z^3 - 30z^2 + 84z + 29).$$

Here, for the zeros of $\rho(z)$ and the growth parameters of the (ρ, σ) procedure, we get

$$z_1 = 1, \quad z_2 = \bar{z}_4 = \exp(i \cdot \frac{\pi}{3}), \quad z_3 = -1,$$

$$\lambda_1 = 1, \quad \lambda_2 = \lambda_4 = \frac{1}{270}, \quad \lambda_3 = -\frac{49}{135}.$$

Subsequent calculation then yields

$$\rho^*(z) = 4z^4 - 7z^3 + 3z^2 + z - 1$$

$$\sigma^*(z) = \frac{1}{360}(521z^4 + 820z^3 - 1344z^2 + 648z + 71).$$

The (R, S) procedure is thus defined by the polynomials

$$R(w,hL) = w^4(1+2hL) - w^3(1+\frac{7hL}{3}) + \frac{3hL}{3} \cdot w^2 + w(1+\frac{hL}{2}) - 1 - \frac{hL}{2}$$

and

$$S(w,hL) = \frac{1}{720}\left[(232 + 521.hL)w^4 + (752 + 820hL)w^3 - (520 + 1344hL)w^2 + \right.$$
$$\left. + (752 + 648.hL)w + 232 + 71hL\right].$$

For the error constants, we find

$$C = -\frac{37}{7560} \quad \text{and} \quad C^* = -\frac{193}{850}.$$

/70

6.3. A special optimal procedure with k = 6.

We choose as the first characteristic polynomial

$$\rho(z) = z^6 - 2z^4 + 2z^2 - 1.$$

This polynomial has the zeros

$$z_1 = 1, \quad z_2 = \bar{z}_3 = \exp(i\,\frac{\pi}{6}), \quad z_5 = \bar{z}_5 = \exp(i\,\frac{5\pi}{6}), \quad z_4 = -1.$$

As the second characteristic polynomial we obtain, in accordance with (1.3.8),

$$\sigma(z) = \frac{1}{3750}(1007.z^6 + 5076z^5 - 2377z^4 - 2033z^3 - 2277z^2 + 5076z + 1007).$$

The growth parameters of the (ρ, σ) procedure are

$$\lambda_1 = 1, \quad \lambda_2 = \lambda_3 = \frac{6951 - 5044.\sqrt{3}}{22500},$$

$$\lambda_4 = -\frac{337}{169}, \quad \lambda_5 = \lambda_6 = \frac{6982 + 5044.\sqrt{3}}{22500}.$$

We then find in the usual way,

$$\varphi^*(z) = 6z^6 - 6z^5 - 6z^4 + 6z^3 + 4z^2 - 4z$$

$$\sigma^*(z) = \frac{1}{360000}(55719z^6 + 202200z^5 - 240031z^4 - 70016z^3 +$$

$$+ 60420z^2 + 62700z - 3551).$$

The characteristic polynomials of the modified procedure are then

$$R(w,hL) = w^6(1 + 3hL) - 3hL.w^5 - 3w^4(1 + 2hL) + 4hL.w^3 + 3w^2(1 + hL) - 2hL.w - 1$$

and

$$S(w,hL) = \frac{1}{360000}\Big[(17552 + 55719hL)w^6 + (95016 + 202200hL)w^5 - (30432 + \quad /71$$

$$+240031.hL)w^4 - (32512 + 70016.hL)w^3 + (-30432 + 60420hL)w^2 +$$

$$+ (95016 + 62700hL)w + 17552 - 3551.hL\Big].$$

As error constants, we have

$$c = -\frac{663}{220500}$$

and

$$c^* = -\frac{4033L}{2410L0}.$$

CHAPTER 7

MODIFIED EXPLICIT PROCEDURES FOR k = 2, k = 4 AND k = 8

7.1. The modified procedure of Nyström (q = 0).

This procedure, often called the midpoint rule, is defined by the polynomial

$$\rho(z) = z^2-1 \quad \text{and} \quad \sigma(z) = 2z$$

The zeros of $\rho(z)$ and the growth parameters of the procedure are

$$z_1 = 1, \quad z_2 = -1, \quad \lambda_1 = 1, \quad \lambda_2 = -1.$$

Since we find for this case

$$\rho^{\circ}(z) = 2z^2 - 2z \quad \text{and} \quad \sigma^{\circ}(z) = 3z-1$$

the modified procedure has the characteristic polynomials

$$R(w,hL) = w^2 . (1 + hL) - hL. w - 1$$

$$S(w,hL) = \frac{1}{2}\left[(4 + 3hL) . w - hL\right].$$

As was shown in Chapter 5, the error constant of the (R, S) procedure has the value

$$C = \frac{1}{6}$$

which is independent of L, and it is equal to the error constant of the midpoint procedure.

7.2. Explicit procedures with k = 4.

Consider those procedures that have the same first characteristic polynomials as the optimal procedures investigated in section 6.2.

I.

$$\rho(z) = z^4 - 1$$

$$\sigma(z) = \frac{4}{3}(2z^3 - z^2 + 2z).$$

The four growth parameters are /73

$$\lambda_1 = 1, \quad \lambda_2 = \lambda_4 = \frac{1}{3}, \quad \lambda_3 = -\frac{3}{5}.$$

Since the supplementary polynomials have the forms

$$\rho^*(z) = 4z^4 - 4z^3$$
$$\sigma^*(z) = \frac{1}{3}(29z^3 - 34z^2 + 23z - 8)$$

it follows that

$$R(w,hL) = w^4(1 + 2hL) - 2hL.w^3 - 1$$
$$S(w,hL) = \frac{1}{6}\left[(16 + 29hL)w^3 - (8 + 34hL)w^2 + (16 + 23hL)w - 6hL\right].$$

The error constant is

$$C = \frac{103}{1440}.$$

II.

$$\rho(z) = z^4 + z^3 - z - 1$$
$$\sigma(z) = 3z^3 + 3z.$$

The growth parameters have the values

$$\lambda_1 = 1, \quad \lambda_2 = \lambda_4 = 1, \quad \lambda_3 = -3.$$

With the supplementary polynomials

$$\rho^*(z) = 4z^4 - z^3 - 3z^2 - z - 1$$

and

$$\sigma^*(z) = \frac{1}{4}(41z^3 - 29z^2 + 19z - 7)$$

we find for the characteristic polynomials of the modified procedure

$$R(w,hL) = w^4(1 + 2hL) + w^3\left(1 - \frac{hL}{2}\right) - \frac{3hL}{2}.w^2 - w\left(1 + \frac{hL}{2}\right) - 1 + \frac{hL}{2}$$
$$S(w,hL) = \frac{1}{8}\left[(24 + 41hL)w^3 - 29hL.w^2 + (24 + 19hL)w - 7hL\right].$$

The error constant is /74

$$C = \frac{1}{20}.$$

III.

$$\rho(z) = z^4 - z^3 + z - 1$$
$$\sigma(z) = \frac{1}{3}(7z^3 - 8z^2 + 7z).$$

For the growth parameters we find

$$\lambda_1 = 1, \quad \lambda_2 = \lambda_4 = \frac{1}{6}, \quad \lambda_3 = -\frac{11}{6}.$$

With

$$\rho^*(z) = 4z^4 - 7z^3 + 3z^2 + z - 1$$

and

$$\sigma^*(z) = \frac{1}{12}(97z^3 - 140z^2 + 91z - 15)$$

we obtain the modified characteristic polynomials

$$R(w,hL) = w^4(1 + 2hL) - w^3(1 + \frac{7hL}{2}) + \frac{3hL}{2}w^2 + w(1 + \frac{hL}{2}) - 1 - \frac{hL}{2}$$

$$S(w,hL) = \frac{1}{24}\left[(56 + 97hL)w^3 - (64 + 140hL)w^2 + (56 + 91.hl)w - 15hL\right].$$

The error constant takes the value

$$C = \frac{29}{160}$$

7.3. An explicit procedure with k = 6.

As in section 6.3, we choose

$$\rho(z) = z^6 - 3z^4 + 2z^2 - 1.$$

Here, for the corresponding $\sigma(z)$ we find

$$\sigma(z) = \frac{1}{20}(209z^5 - 440z^4 + 474z^3 - 440z^2 + 209z).$$

From this we get the growth parameters

/75

$$\lambda_1 = 1, \quad \lambda_2 = \lambda_6 = \frac{669 - 536\sqrt{3}}{270}, \quad \lambda_3 = \lambda_5 = \frac{669 + 536\sqrt{3}}{270},$$

$$\lambda_4 = -\frac{401}{45}.$$

With the polynomials

$$\rho^*(z) = 6z^6 - 6z^5 - 8z^4 + 8z^3 + 4z^2 - 4z$$

$$\sigma^*(z) = \frac{1}{720}(12917z^5 - 26163z^4 + 25342z^3 - 10818z^2 + 9597z - 1435)$$

the modified characteristic polynomials are written

$$R(w,hL) = w^6(1 + 3hL) - 3hL.w^5 - w^4(2 + 4hL) + 4hL.w^3 + w^2(2 + 2hL) -$$

$$-2hL.w - 1$$

and

$$S(w,hL) = \frac{1}{1440}\left[(4704 + 12017.hL)w^5 - (7136 + 26103hL)w^4 + \right.$$

$$+ (7504 + 25342hL)w^3 - (7136 + 16818hL)w^2 +$$

$$\left. + (4704 + 9597hL)w - 1436hL\right].$$

The corresponding error constant has the value C = 1097/7560.

CHAPTER 8

/76

EXAMPLES

8.1. Discussion of nonlinear system.

Consider the following system of four nonlinear first-order differential equations:

$$\begin{aligned} x' &= u \quad , & x(0) &= 1 \\ u' &= -x \cdot r^{-3} \quad , & u(0) &= 0 \\ y' &= v \quad , & y(0) &= 0 \\ v' &= -y \cdot r^{-3} \quad , & v(0) &= 1 \, . \end{aligned} \tag{8.1.1}$$

with $r^2 = x^2 + y^2$, $' = \frac{d}{dt}$.

This system, which one encounters in connection with the two-body problem, is discussed in the appendix of [22]. It has the exact solution

$$X(t) = \begin{pmatrix} x(t) \\ u(t) \\ y(t) \\ v(t) \end{pmatrix} = \begin{pmatrix} \cos t \\ -\sin t \\ \sin t \\ \cos t \end{pmatrix}$$

The corresponding Jacobian matrix has the form

$$F_X(t) = F_X(t, X(t)) = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{3}{2}\cos(2t) + \frac{1}{2} & 0 & \frac{1}{2}\sin(2t) & 0 \\ 0 & 0 & 0 & 1 \\ \frac{3}{2}\sin(2t) & 0 & -\frac{3}{2}\cos(2t) + \frac{1}{2} & 0 \end{pmatrix} .$$

If we use the Euclidean vector norm, we get, in accordance with section 4.2,

$$\mu\left[F_X(t)\right] = \text{greatest eigenvalue of the matrix } \frac{1}{2}\left(F_X^{\circ}(t) + F_X(t)\right).$$

In our case, this matrix is symmetric, and a simple calculation yields, for the eigenvalues,

/77

$$\Lambda_1 = \Lambda_2 = 0 \, , \quad \Lambda_3 = \frac{3}{2} \, , \quad \Lambda_4 = -\frac{3}{2} \, .$$

In the notation of our section 4.2, we have

$$\mu\left[F_X(t)\right] = \frac{3}{2} \, , \quad \mu'\left[F_X(t)\right] = -\frac{3}{2} \, .$$

In connection with section 4.3, we then have

$$U_3 = 3 \cdot \max_{\substack{\{\mu \mid \lambda_\mu > 0\} \\ \mu \neq 1}} \frac{\lambda_\mu - 1}{1-\cos\varphi_\mu}, \quad U_4 = 3 \cdot \max_{\{\mu \mid \lambda_\mu < 0\}} \frac{-(\lambda_\mu - 1)}{1-\cos\varphi_\mu},$$

$$V_3 = 3 \cdot \max_{\substack{\{\mu \mid \lambda_\mu > 0\} \\ \mu \neq 1}} \frac{\lambda_\mu}{1-\cos\varphi_\mu}, \quad V_4 = 3 \cdot \max_{\{\mu \mid \lambda_\mu < 0\}} \frac{-\lambda_\mu}{1-\cos\varphi_\mu}.$$

If we now apply the modified Milne-Simpson procedure for numerical integration, we find the following stabilization parameters: For relative stability

$$L > 2$$

For absolute stability,

$$L > \frac{1}{4}.$$

8.2. A numerical illustration.

To compare the effectiveness of the modified difference procedure with that of the classical procedure, we have integrated the following two simple initial-value problems:

$$x' = A - Ax, \quad x(0) = 0, \quad A = 10, \tag{8.2.1}$$

and

$$x' = At - At.x, \quad x(0) = 0, \quad A = 2. \tag{8.2.2}$$

In both cases, the exact solution can be given explicity. It is

$$x(t) = 1 - \exp(10t) \tag{8.2.3}$$

for the problem (8.2.1) and /78

$$x(t) = 1 - \exp(-t^2) \tag{8.2.4}$$

for the problem (8.2.2). Both solutions exhibit the same behavior: they rapidly approach the value 1 monotonically with increasing values of t. For the first initial-value problem, the interval of integration [0, 6] was chosen whereas the problem (8.2.2) was integrated over the interval [0, 3]. The initial values were "exact", namely, equal to the exact solution of the corresponding initial-value problem at the points $t_0 = 0$ and $t_1 = h$ in the first case and $t_0 = 0$, $t_1 = h$, $t_2 = 2h$, $t_3 = 3h$ in the other.

Example 1: Numerical integration of (8.2.1).

Classical optimal procedure (Milne-Simpson), $\rho(z) = z^2 - 1$

Modified procedure, $R(w,hL) = w^2(1 + hL) - hL \cdot w - 1.$

Zeros of $\rho(z)$: $z_1 = 1, \quad z_2 = -1$.

Growth parameters of the (ρ, σ) procedure: (ρ, σ): $\lambda_1 = 1, \quad \lambda_2 = -\frac{1}{3}$.

Stabilization inequalities:

for relative stability, $L > \frac{40}{3}$ (in accordance with (3.1.8)),

for absolute stability, $L > \frac{10}{3}$ (according to 3.1.9)).

Step: $h = 0.1$.

Numerical values: /79

t_n	$L = 0$		$L = 9$	
	x_n	e_n	x_n	$10^8 \cdot e_n$
.				
.				
.				
2.0	1.5479	0.5479	0.999997	-1254
2.1	0.2515	-0.7485	1.000009	992
2.2	2.0225	1.0225	0.999993	-738
2.3	-0.3967	-1.3967	1.000006	568
.				
.				
.				
5.0	$6.3 \cdot 10^3$	$6.3 \cdot 10^3$	1.000009	-0.44
5.1	$-8.7 \cdot 10^3$	$-8.7 \cdot 10^3$	1.000000	0.33
		etc.		

Example 2. Numerical integration of (8.2.1).

Classical optimal procedure, $\rho(z) = z^4 - 1$.

Modified procedure, $R(w,hL) = w^4(1 + 2hL) - 2hL \cdot w^3 - 1$.

Zeros of $\rho(z)$, $z_1 = 1, \; z_2 = \bar{z}_4, \; \exp(i\frac{\pi}{2}), \; z_3 = -1.$

Growth parameters of (ρ, σ), $\lambda_1 = 1, \; \lambda_2 = \lambda_4 = \frac{1}{45}, \; \lambda_3 = -\frac{19}{45}$.

Stabilization inequalities:

for relative stability, $L > \frac{173}{9}$

for absolute stability, $L > \frac{33}{9}$.

Step, h = 0.05.

Numerical values: /80

	L = 0		L = 10	
t_n	x_n	c_n	x_n	$10^3 . c_n$
.				
.				
.				
2.00	1.0009	0.0009	1.000000	-227
2.05	0.9879	-0.0121	1.000000	200
2.10	1.0149	0.0149	1.000000	-170
2.15	0.9817	-0.0183	1.000000	167
.				
.				
.				
5.00	$2.3 . 10^3$	$2.3 . 10^3$	1.000000	-0.99
5.05	$-2.3 . 10^3$	$-2.3 . 10^3$	1.000000	0.87
		etc.		

Example 3. Numerical integration of (8.2.2).

Classical optimal procedure, as in Example 1.

Modified procedure, as in Example 1.

Stabilization inequalities:

for relative stability, $L > 4$,

for absolute stability, $L > 1$.

Step, h = 0.1.

Numerical values: /81

t_n	L = 0		L = 10	
	x_n	e_n	x_n	e_n
.				
.				
.				
1.0	0.671531	0.039411	0.634500	0.002406
1.1	0.669709	-0.032004	0.703746	0.001943
1.2	0.805822	0.042750	0.764658	0.001503
.				
.				
.				
2.7	0.736559	-0.262758	0.999021	0.000003
2.8	1.313324	0.313713	0.999008	0.000001
2.9	0.622930	-0.376847	0.999778	0.000001
3.0	1.455365	0.455469	0.999977	0.000000

Example 4. Numerical integration of (8.2.2).

Same difference procedure in Example 3. Here, h = 0.05 was chosen for the step.

Numerical values:

t_n	L = 0		L = 10	
	x_n	e_n	x_n	e_n
.				
.				
.				
1.00	0.642159	0.010039	0.632040	0.000020
1.05	0.660235	-0.007725	0.652785	0.000015
1.10	0.712207	0.010404	0.702548	0.000745
.				
.				
.				
2.85	0.909078	-0.091025	0.999701	0.000001
2.90	1.100530	0.100753	0.999778	0.000001
2.95	0.888857	-0.110833	0.999991	0.000000
3.00	1.122295	0.122418	0.999977	0.000000

8.3. Remarks. /82

The above-given results indicate clearly the completely different error behavior of the solutions of the classical and the modified difference procedures.

If we again look at the stabilization inequalities derived in Chapters 3 and 4, we see that the right-hand members of these relationships are independent of the step h (since we have investigated the asymptotic behavior of the approximate solution and of the discretization error). However, these bounds for "large" values of h probably depend on the chosen step of the difference procedure. This

became obvious with the numerical integration. Theoretically, conditional stability should no longer be present in the first example for $L > 10/3$ or in the second example for $L > 38/9$. However, it was shown that the corresponding solution was stable in the first example only for $h = 0.1$ and $L = 4$ whereas this solution indicated absolutely stable behavior for $h = 0.05$ and for $L = 4$. This showed up even more clearly in the second example. There, the solution exhibited conditional stability for all $L \leq 9$ and $h = 0.1$. In contrast, the behavior with step $h = 0.05$ was absolutely stable for $L = 6$. Similar conclusions were drawn in the integration of (8.2.2). These remarks bring up another problem: What choice of hL yields an optimal error behavior for a given initial-value problem, and can the step h be written in advance? The answers to these questions are not yet known.

The numerical calculations were made on the control data 1604A of the computational center of the ETH (Eidgenössische Technische Hochschule) in Zürich and on the IBM 360/50 at Dalhousie University in Halifax, Canada.

REFERENCES / 83

The following bibliographic list contains, in addition to books and articles that were used directly, other publications (in particular new data) that touch the set of problems treated in the present dissertation but were not decisive for its presentation. Such titles are denoted by an asterisk.

1. Ahlfors, L. V. Complex Analysis. McGraw-Hill, New York (2nd ed.), 1966.
2. Bellman, R. Introduction to Matrix Analysis, McGraw-Hill, 1960.
3.* Brown, R., J. Riley, and M. Bennet Stability Properities of Adam-Moulton Type Methods. Math. Comp., 19, 1965.
4.* Butcher, J. C. A Modified Multistep Method for the Numerical Integration of Ordianry Differential Equations. J. ACM, 12, 1965.
5.* Case, J. H. A Note on the Stability of Predictor-Corrector Techniques. Manuscript, 1964.
6. Cesari, L. Asymptotic Behaviour and Stability Problems in Ordinary Differential Equations. Academic Press, New York, 1963.
7.* Chase, P. Stability Properties of Predictor-Corrector Methods for Ordinary Differential Equations. J. ACM, 9, 1962.
8.* Crane, R. L. and R. W. Klopfenstein A Predictor-Corrector Algorithm with Increase Range of Absolute Stability. J. ACM, 12, 1965.
9.* Crane, R. L. and R. J. Lambert Stability of Generalized Corrector Formulas, J. ACM, 9, 1962
10. Dahlquist, G. Convergence and Stability in the Numerical Integration of Ordinary Differential Equations. Math. Scand., 4, 1956.
11. Dahlquist, G. Stability and Error Bounds in the Numerical Integration of Ordinary differential Equations. Trans. Royal Inst. Technol., Stockhlom, No. 130, 1959.
12. Dahlquist, G. Stability Questions for Some Numerical Methods for Ordinary Differential Equations. Proc. Symposia Appl. Math., 15, 1963.
13. Dahlquist, G. A Special Stability Problem for Linear Multistep Methods. Nord. Tidskr. Inform. Beh., 3, 1963.
14.* Emanuel, G. The Wilf Stability Criterion for Numerical Integration. J. ACM, 10, 1963.

15.* Fehlberg, E. Numerisch stabile Interpolationsforemeln mit günstiger Fehlerfortpflanzung für Differentialgleichungen erster und zweiter Ordnung. / 84
Z. angew. Math. Mech., 41, 1961.
16. Goldberg, S. Introduction to Difference Equations. Wiley, New York, 1961.
17.* Gragg, W. B. and H. J. Stetter Generalized Multistep Predictor-Corrector Methods, J. ACM, 11, 1964.
18. Gröbner, W. Matrizenrechnung, B. I. Hochschultaschenbücher. Mannheim, No. 103, 1966.
19. Hamming R. W. Stable Predictor-Corrector Methods for Ordinary Differential Equations. J. ACM, 6, 1959.
20. Henrici P. Discrete Variable Methods in Ordinary Differential Equations. Wiley, New York, 1962.
21. Henrici, P. Problems of Stability and Error Propagation in the Numerical Integration of Ordinary Differential Equations. Proc. Int. Congr. Math., 1962.
22. Henrici, P. Error Propagation for Difference Methods. Wiley, New York, 1963.
23. Henrici, P. The Propagation of Error in the Digital Integration of Ordinary Differential Equations. Manuscript, 1964.
24. Hochstadt, H. Differential Equations. Holt, Reinhart and Winston, New York, 1965.
25.* Hull, T. E. and W. A. J. Luxemburg Numerical Methods and Existence Theorems for Ordianry Differential Equations. Num. Math., 2, 1960.
26.* Hull, T. E. and A. C. R. Newbery Integration Procedures which minimize propagated Errors. J. Soc. Indust. Appl. Math., 9, 1961.
27. Hull, T. E. and A. C. R. Newbery Corrector Formulas for Multistep Integrations Methods. J. Soc. Indust. Appl. Math., 10, 1962.
28.* Hull, T. E. A Search for Optimum Methods for the Numerical Integration of Ordinary Differential Equations. SIAM Rev., 9, 1967.
29.* Krogh, F. T. Predictor-Corrector Methods of High Order with Improved Stability Characteristics, J. ACM, 13, 1966.
30.* Lambert, J. D. and B. Shaw A Generalization of Multistep Methods for Ordinary Differential Equations, Num. Math., 8, 1966.
31.* Lambert, R. J. An Analysis of the Numerical Stability of Predictor-Corrector Solutions of Nonlinear Ordinary Differential Equations, SIAM J. Num. Anal. 4, 1967.
32.* Lambert, R. J. and A. R. Mitchell On the Solution of $y' = f(x, y)$ by a Class of / 85
High Accuracy Difference Formulae of Low Order. Z. angew. Math. Phys. 13, 1962.
33. Marden, M. Geometry of Polynomials, Am. Math. Soc., Providence, (2nd ed.) 1966.
34. Milne, W. E. Numerical Solution of Ordinary Differential Equations, Wiley, New York, 1953.
35.* Milne, W. E. and R. R. Reynolds Stability of a Numerical Solution of Differential Equations, J. ACM, 6, 1959, (part I), J. ACM, 7, (part II), 1960.
36. Obreschkov, N. Verteilung und Berechnung der Nullstellen reeller Polynome, VEB Deutscher Verl. d. Wissensch., Berlin, 1963.
37.* Ralston, A. Relative Stebility in the Numerical Solution of Ordinary Differential Equations. SIAM Rev., 7, 1965.
38.* Reimer, M Eine Fehlerabschätzung für lineare Differenzenverfahren, Num. Math., 7, 1955.

39. Rutishauser, H. Ueber die Instabilitat von Methoden zur Integration gewöhnlicher Differentialgleichungen, Z. angew. Math. Phys., 3, 1952.
40. Silverman, R.A. Introductory Complex Analysis, Prentice-Hall, Englewood Cliffs, N.J., 1967.
41.* Stetter, H.J. A Study of Strong and Weak Stability in Discretization Algorithms, SIAM J. Num. Anal., 2, 1965.
42.* Stetter, H.J. Stabilizing Predictors for Weakly Unstable Correctors, Math. Compl., 9, 1965.
43.* Urabe, M. Theory of Errors in Numerical Integration of Ordinary Differential Equations, J. Sci. Hiroshima Univ., Ser. A-1, 25, 1961.
44.* Wilf, H.S. A Stability Criterion for Numerical Integration, J. ACM, 6, 1959.
45.* Wilf, H.S. Maximally Stable Numerical Integration, SIAM J., 8, 1960.
46. Wilkinson, J.H. Rounding Errors in Algebraic Processes, Prentice-Hall, Englewood Cliffs, N.J., 1963.
47. Wilkinson, J.H. The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1965.

BIOGRAPHICAL SKETCH

I was born in [redacted] and am a citizen of Hemburg, St. Gallen. I went to primary school in the years 1948-1954 in Netstal. After finishing the secondary school there, in the spring of 1957 I took my examination for the scientific branch of the Glarus Canton School. In autumn 1960, I obtained the confederational (Eidgenössisch) certificate of qualification (type C). Then, I began the study of mathematics at division IX of the confederational technical high school in Zürich. In the autumn of 1964, I obtained the diploma as mathematician. My thesis (Stable methods of maximum order for numerical solution of ordinary differential equations) was under the direction of Professor P. Henrici. In the same year, I obtained from the technical high school the capability certificate for teaching at an advanced level. Toward the end of 1964 I began my dissertation, again under the direction of Professor Henrici. After serving as auxiliary teacher at the Canton School of Zürcher Oberland at Wetzikon since the autumn of 1963, I was chosen in April 1965 as chief mathematics teacher at that school. Since September 1967 I have been living in Halifax, Canada, where I work as assistant professor of mathematics at Dalhousie University.

Translated for the National Aeronautics and Space Administration by Scripta Technica, Inc. NASW-2036.